Author:
Cooper, Philip J

Title:
The Modified 5-Year-Olds' Index – testing reliability and the impact of training level

The Modified 5-Year-Olds' Index – testing reliability and the impact of training level

Philip James Cooper

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Dental Surgery by advanced study in Orthodontics in the Faculty of Health Sciences

Bristol Dental School

17th July 2020

Word count: 19,861

ABSTRACT

Aim: To test the reliability of the Modified 5-Year-Olds' Index using assessors with a range of experience, and to determine whether calibration improves reliability.

Design: Prospective reliability study.

Setting: Bristol Dental School, University of Bristol.

Methods: Fifty study models of non-syndromic 5-year-olds with a repaired UCLP were selected from the CCUK archives. Fifteen participants with a range of clinical experience in orthodontics were divided into three groups of equal experience. Each group was given differing amounts of information: Group 1 information sheet; Group 2 information sheet and reference models; Group 3 calibration course, information sheet and reference models. Each participant scored the 50 models using the Modified 5-Year-Olds' Index on two occasions at least four weeks apart. ICCs calculated from a two-way random effects model were used to calculate intra-rater reliability and inter-rater reliability comparing assessors' scoring to the experts' consensus scores (gold standard).

Results: Group 2 (ICC 0.80 – 0.93) and Group 3 (ICC 0.80 – 0.91) demonstrated high levels of intra-rater agreement, with lower levels shown by Group 1 (0.68 – 0.93). Inter-rater agreement was high in Group 2 (ICC 0.87 – 0.93) and Group 3 (0.82 – 0.91), with Group 1 showing the lowest levels of agreement (ICC 0.69 – 0.94). The level of training of the assessors in Group 1 appears to influence reliability scores, with high intra- and inter-rater reliability scores of the consultant and post-CCST trainee at a similar level to those of the same grade in Groups 2 and 3.

Conclusions: The Modified 5-Year-Olds' Index is a reliable method of assessing outcomes when model scoring is carried out by consultants or post-CCST trainees. Calibration in use of the Index does not improve reliability.

DEDICATION AND ACKNOWLEDGEMENTS

With sincere thanks to my academic supervisors Tony Ireland, Nikki Atack, Sam Leary and Jonathan Sandy for their considerable guidance, advice and patience over the course of this research project. Many thanks to those who gave up their time to participate in the project – Consultant Orthodontists Christian Day, Tim Jones and Julie Williams, post-CCST orthodontic specialty trainees Sean Hamilton, Tara Lee and Graham Oliver and pre-CCST orthodontic specialty trainees Zainab Al Saffar, Aliaa Fauzi, Kyle Durman, Jen Jopson, Saleem Hasanally, Jennifer Haworth, Nadine Homoud, Charlotte Molyneaux and Miesha Virdi.

Thanks also to my clinical supervisors who have been instrumental in providing high quality orthodontic training – Consultants Nikki Atack, Tony Ireland, Tim Jones, Kate House, Farnaz Parvizi and Julie Williams, and to my very supportive Educational Supervisor Christian Day.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: DATE:

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY OF TERMS

BCLP	Bilateral cleft lip and palate
CAPS-A	Cleft audit protocol for speech – Augmented
CCST	Certificate of Completion of Specialty Training
CCUK	Cleft Care UK
CI	Confidence Interval
CIG	Cleft Implementation Group
CLP	Cleft lip and palate
CP	Cleft palate
CRANE	Cleft Registry and Audit Network
CSAG	Clinical Standards Advisory Committee Group
DMFT	Decayed Missing Filled Teeth
EUROCRAN	European Collaboration on Craniofacial Anomalies
GOLSON Yardstick	Great Ormond Street London and Oslo Yardstick
GOS.SP.ASS	Greater Ormond Street Speech Assessment
HB	Huddart and Bodenham
HRQOL	Health related quality of life
ICC	Intraclass Correlation Coefficient
IOFTN	Index of Orthognathic Functional Treatment Need
IOTN	Index of Orthodontic Treatment Need
k	Number of raters/measurements
MDT	Multidisciplinary Team
MHB	Modified Huddart and Bodenham

NHS	National Health Service
OJ	Overjet
PAR	Peer Assessment Rating
PROMs	Patient Reported Outcome Measures
QOL	Quality of life
REC	Research Ethics Committee
SWAG	Standard Way to Assess Grafts
UCLP	Unilateral cleft lip and palate
VAS	Visual analogue scale
VLS	Vermillion, lip, scar

1.0 INTRODUCTION

Evaluation and improvement of quality of care provided to patients are essential in clinical practice. In 1992 an international multicentre study on the outcomes of cleft care across six cleft centres in Europe demonstrated that UK units performed relatively badly in comparison to those in mainland Europe (Shaw *et al.*, 1992b). The UK Clinical Standards Advisory Committee Group (CSAG) conducted an enquiry (Sandy *et al.*, 1998) and subsequently made a number of recommendations that were accepted by the UK government. Fifteen years after the CSAG recommendations, the Cleft Care UK study examined the impact of centralisation on cleft services. An improvement in dentoalveolar outcomes was found at the age of 5 years using the 5-Year-Olds' Index (Al-Ghatam *et al.*, 2015).

As outcomes in cleft care have improved, it has become increasingly difficult to discriminate between the outcome categories of the 5-Year-Olds' Index, with a higher proportion of study models scored in the better categories of the index. It is therefore difficult to demonstrate continued improvement in outcomes for the purposes of audit. The Modified 5-Year-Olds' Index was developed to address this problem, and was found to be both reliable and able to discriminate more sensitively within the good outcome categories than the 5-Year-Olds' Index (Mittal *et al.*, 2018).

Although the Modified 5-Year-Olds' Index was found to be reliable, testing was carried out using experienced examiners who also helped develop the index. This dissertation aims to assess both the effect of experience and whether calibration and use of a set of reference models is required to use the index reliably. This will provide evidence to support use of the

Modified 5-Year-Olds' Index as an audit tool in measuring the outcome of primary cleft surgery in individuals born with unilateral cleft lip and palate (UCLP).

2.0 REVIEW OF THE LITERATURE

2.1 Cleft lip and palate

Cleft lip and/or palate (CLP) is a common facial birth defect, affecting approximately 1 in 700 live births (Mossey *et al.*, 2009). There is wide variation in prevalence across racial and ethnic groups, ranging from the highest reported rates of 1 in 500 live births in Asian and Native American populations, to 1 in 2500 in African populations (Dixon *et al.*, 2011). Isolated cleft palate occurs in approximately 1 in 2000 live births (Gorlin *et al.*, 2001).

Cleft lip with or without cleft palate occurs more frequently in males, whilst there is an increased frequency of isolated cleft palate in females (Mossey *et al.*, 2009). Cleft lip with or without cleft palate and isolated cleft palate are often associated with other significant congenital anomalies, with approximately 30% of cases of CLP associated with defects including recognised syndromes and chromosomal anomalies (Calzolari *et al.*, 2007). There are over 300 known syndromes that have clefting of the lip or palate as an associated feature, including syndromes such as Pierre Robin sequence, ectodermal dysplasia and Treacher Collins amongst others (Akram *et al.*, 2015). Around 70% of cases occur as an isolated cleft of the lip and/or palate.

The Cleft Registry and Audit Network (CRANE) database was established in 2000 by the UK Department of Health to collect information on all children born with cleft lip and/or cleft palate in England, Wales and Northern Ireland. The database collects information on birth, demographics and cleft, and also records information on cleft-related treatment and outcomes. 20,013 children were born and registered on the database between 1st January

2000 and 31st December 2018 (Medina *et al.*, 2019). In 2018, cleft palate without lip involvement remained the most common of the cleft types, with 39% of cases recorded.

2.1.1 Aetiology

The precise aetiology of non-syndromic CLP is unknown, but is thought to be multifactorial involving both genetic and environmental factors (Schutte and Murray, 1999). The sibling risk for CLP is approximately 30 times that of the normal population (Akram *et al.*, 2015). However, the importance of environmental factors is highlighted by the concordance rate of only 25-45% in identical twins and 3-6% in non-identical twins (Mitchell and Risch, 1992).

Various environmental risk factors have been linked to CLP. Maternal smoking during pregnancy is associated with increased prevalence of CLP and isolated cleft palate (Little *et al.*, 2008). The role of maternal alcohol consumption during pregnancy is positively associated with orofacial clefts in some studies (Munger *et al.*, 1996, Shaw and Lammer, 1999, Werler *et al.*, 1991), but not in others (Meyer *et al.*, 2003). Studies on maternal nutrition find that it may play a role in orofacial clefts. A meta-analysis published in 2008 suggests that multivitamin intake during early pregnancy may afford some protection against oral clefts, although most of the evidence is from observational studies (Johnson and Little, 2008).

2.1.2 Pathophysiology

Development of the lip and palate usually occurs between the 4th and 11th weeks *in utero* and involves a complex series of events, including cell migration, growth, differentiation and

apoptosis. During facial morphogenesis, neural crest cells migrate into the oro-facial region, where they form all skeletal, connective, and dental tissues, apart from enamel.

The neural crest cells form five distinct mesenchymal processes: one frontonasal, two maxillary and two mandibular processes. The mandibular processes join in the midline, forming the lower jaw and lower lip. The palate develops from fusion of the nasal and maxillary processes in a complex process. Mesenchymal cells grow from the maxillary processes to project vertically on both sides of the developing tongue, which then elevate towards each other and join to create the secondary palate. These then form the primary palate, upper lip and nose by fusing anteriorly with the lateral and medial nasal processes. Cleft lip and/or palate is a result of failure of fusion of any of the maxillary or nasal processes.

2.1.3 Classification

There are a number of classifications used to describe the different types of cleft lip and/or palate, most of which are based upon the processes of facial embryology.

In the UK the LAHSAL system, as shown in Figure 1, is used to classify clefts and to document them on the CRANE database. The LAHSAL classification is a modified version of the LAHSHAL system first described by Otto Kriens (1989). Modification by omitting one 'H' was recommended by the Royal College of Surgeons of England in 1995 in order to simplify the classification (Hodgkinson *et al.*, 2005). The classification splits the mouth into six parts:

- Right lip
- Right alveolus

- Hard palate
- Soft palate
- Left alveolus
- Left lip

The code indicates for each of the six sections whether there is a complete cleft (upper case letter), incomplete cleft (lower case letter), or no cleft. The classification has been tested and found to have a high degree of intra- and inter-observer reliability, with Kappa statistics ranging from 0.809 to 0.992 (McBride *et al.*, 2013).

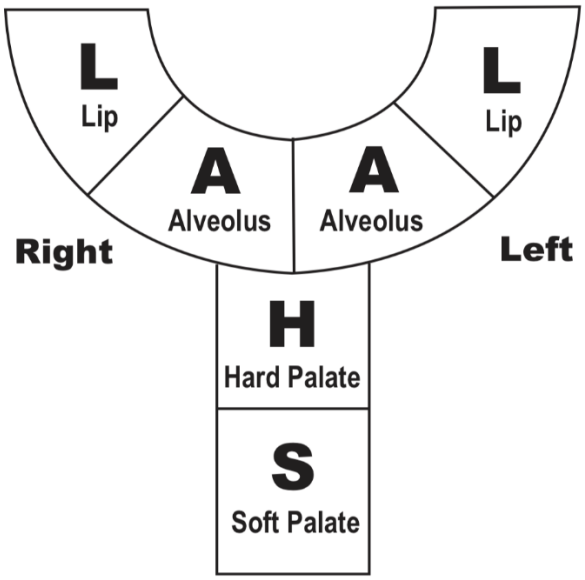


Figure 1: LAHSAL Classification (reproduced from Hodgkinson *et al.,* 2005)

2.1.4 Diagnosis & Treatment of CLP

Children born with CLP require complex long-term treatment and follow up by a multidisciplinary team. In the UK, expectant parents who have received a pre-natal diagnosis of a cleft, or babies with a previously undiagnosed cleft which is identified at birth,

should be referred to their local cleft team within 24 hours of diagnosis. It is the NHS Cleft Lip and Palate Service which provides care and support to both cleft affected children and their families, from diagnosis to adulthood. This multidisciplinary team comprises clinicians from a number of disciplines and includes cleft surgeons, cleft nurses, orthodontists, ENT surgeons, paediatric dentists, paediatricians, speech and language therapists, psychologists and geneticists. The timeline for diagnosis and treatment of CLP is illustrated in Table 1 and will now be described further.

Improved ultrasound scanning has led to an increase in antenatal diagnosis of clefts (Maarse *et al.*, 2010). Early diagnosis gives parents time to prepare emotionally for the birth and to learn more about CLP through prenatal support from the cleft team. However, despite improvements to antenatal scanning, at least 20% of cases of CLP go undiagnosed (Hodgkinson *et al.*, 2005).

The main aims of treatment are to achieve good facial aesthetics, good function of the lip and palate to facilitate normal eating, drinking and speech development, and optimum facial growth and development to prevent or minimise deformity secondary to impaired growth.

For many years, treatment protocols involved presurgical orthopaedics. It was claimed that this treatment could improve arch form, facilitate surgical closure and improve outcomes in terms of aesthetics, feeding and speech. However, this has largely fallen out of favour, with a systematic review of the literature demonstrating no long-term positive effects for almost all treatment outcomes (Uzel and Alparslan, 2011). A recent cohort study demonstrated

significantly worse dental arch and maxillomandibular relationship outcomes in patients treated using presurgical orthopaedics (Kornbluth *et al.*, 2018). However, there is some evidence to demonstrate improvement in nasal symmetry of patients who undergo nasoalveolar moulding, an evolution of the original presurgical orthopaedic technique (Kirbschus *et al.*, 2006, Niranjane *et al.*, 2014).

Age	**Care**
Birth – 8 weeks	• Cleft Nurse Specialist visit within 24 hours of diagnosis • Specialist feeding assessment & management • Meet Cleft MDT • Clinical psychology support offered
3-6 months	• Lip repair
6-12 months	• Palate repair • Paediatric dentist for dental health education/advice
18 months	• Speech and language assessment & management
3-7 years	• Pyschological support prior to school entry • Surgery to revise lip and speech if necessary • Full MDT and records at 5 years
8-14 years	• Assessment by cleft team Orthodontist, Paediatric dentist and surgeon responsible for alveolar bone graft at aged 7-8 years • Paediatric dentistry • Orthodontic treatment • Full MDT & records at 10 years

	• Alveolar bone grafting at the age of 8-10 years, prior to eruption of the permanent canine • Management of speech and hearing problems
15-21+ years	• Orthognathic surgery with associated orthodontics if required • Revision surgery where required • Post-orthognathic surgery records and speech assessment • Speech revision surgery if required • Rhinoplasty

Table 1: Timeline of treatment. Adapted from NHS Standard Contract for Cleft Lip and/or Palate Services 2013 (https://www.england.nhs.uk/wp-content/uploads/2013/06/d07-cleft-lip.pdf)

Primary surgery is carried out in two separate operations. The first is repair of the lip and nasal form, usually carried out between the ages of 3 and 6 months. The second is palate repair aged between 9 and 12 months (Table 1). Early closure of the palate and alveolus is linked with poor maxillary growth due to scarring (Ross, 1970), resulting in a more hypoplastic maxilla, a concave mid-face and deformed dental arch (Shi and Losee, 2014). However, early closure results in improved speech outcomes, and therefore a balance must be struck between closing the palate early for this reason and delaying surgery to maximise maxillary growth potential (Chapman *et al.*, 2008). There is currently no consensus on treatment protocol or optimal timing of cleft palate closure (Kappen *et al.*, 2017).

Speech abnormalities are common in children born with cleft palate which can affect social, educational and childhood development. Speech and language therapy plays a key role in speech development for children born with a cleft. A cleft palate may cause velopharyngeal insufficiency, which often persists after primary surgery. Velopharyngeal insufficiency is a lack of coordination in closure of the soft palate to the posterior pharyngeal wall, leading to problems with speech, eating and breathing. Patients undergo combined assessment by the speech and surgical teams, and may require re-repair of the palate or pharyngoplasty to correct such problems (Tollefson *et al.*, 2012).

Clefting of the alveolus occurs in approximately 75% of children born with a cleft lip or cleft lip and palate (Guo *et al.*, 2011). Failure to correct this can also lead to speech problems, oronasal fistulae, fluid reflux, transverse maxillary deficiency, lack of bone support for anterior teeth, dental crowding and asymmetry (Waite and Waite, 1996). Alveolar clefts are repaired by secondary alveolar bone grafting at approximately 8-10 years of age, prior to eruption of the permanent canine. A period of orthodontic treatment to expand and improve the shape of the dental arch usually precedes the graft. In this way, space is created to facilitate placement of bone and allow eruption of the canine or lateral incisors at the site of the cleft, and hopefully into a normal position. The iliac crest or tibia are commonly used as donor sites to harvest the bone. A Cochrane review by Guo *et al.* (2011) found limited evidence to support using artificial bone over autografting.

Orthodontic treatment is commonly indicated in the permanent dentition to improve or correct a number of problems, including crowding, anterior and posterior crossbites and missing teeth. Patients born with CLP have an increased frequency of dental anomalies such

as hypodontia, supernumeraries, hypoplasia and abnormalities in tooth size and shape (Tan *et al.*, 2018). As a result of impaired maxillary growth, there is sometimes a significant Class III skeletal relationship requiring orthognathic surgery after completion of growth. Secondary rhinoplasty is carried out after facial growth is complete, with placement of cartilage grafts for support and reinforcement of the nasal structure where required (Kaufman *et al.*, 2012).

2.2 Cleft care in the UK

In 1992 a six centre study was conducted to look at the outcomes of cleft care across Europe (Shaw *et al.*, 1992a). The study aimed to identify treatment factors resulting in favourable and unfavourable outcomes in dental arch relationships, facial aesthetics and craniofacial form. From this, the authors aimed to create treatment and follow up protocols in order to standardise care and improve outcomes. They found a wide variation in the timing of treatments, a variety of surgical techniques and great variability in the workload per surgeon. Of the six study units, the two in the UK performed relatively poorly in comparison to those in mainland Europe. They had the greatest variability in treatment protocols and a relatively small number of patients treated by a large number of surgeons. A number of different outcomes measures were used in the study; dental arch relationships were assessed on study models using the GOSLON yardstick (Mars *et al.*, 1992, Mølsted *et al.*, 1992), craniofacial form on cephalometric radiographs (Mars *et al.*, 1992, Mølsted *et al.*, 1992) and nasolabial appearance on clinical photographs (Asher-McDade *et al.*, 1992). Outcomes in the two UK centres were poor compared with those operating with standardised protocols in a centralised service.

Auditing surgeons operating on small numbers of patients is difficult, owing to the time required to generate a sufficiently large sample size for meaningful results (Shaw *et al.*, 1992b). Further studies in the UK were subsequently carried out, investigating organisation of cleft units in the UK. Williams *et al.* (1994) found a third of surgeons performing cleft operations carried out fewer than five primary surgeries per year. Further work revealed a large proportion of surgeons in the UK were classified as low volume operators, performing less than 10 primary surgeries per year (Williams *et al.*, 1996). It was generally felt that change was needed in the UK in order to improve outcomes (Shaw *et al.*, 1996). Areas in need of change that were highlighted included volume of surgery, surgical proficiency, organisation, audit and research.

2.2.1 CSAG report – 1998

With the above mentioned publications finding less than ideal quality of cleft care in the UK, in 1995 the Royal College of Surgeons and the Standing Dental Advisory Committee appealed to the Department of Health to conduct an inquiry into the standards of cleft care in the UK (Hathorn *et al.*, 2006). The Clinical Standards Advisory Committee Group (CSAG) subsequently conducted an enquiry over a period of 15 months starting in 1996 (Sandy *et al.*, 1998).

A research team was appointed, with senior plastic and maxillofacial surgeons appointed as coordinators across the UK. Outcome data was gathered for 5 and 12-year-old children born with non-syndromic unilateral cleft lip and palate (UCLP). Patients with UCLP were chosen, as this encompasses the full range of surgeries, with both lip and palate repair, and alveolar bone grafting. The 5-Year-Olds' Index (Atack *et al.*, 1997a) was used to measure the

outcome of primary surgery through the dental arch relationships at 5 years of age, before results are potentially distorted by the effects of orthodontics or bone grafting. Speech was also assessed in the 5-year-old age group. Assessment of secondary alveolar bone grafting and facial growth could be assessed in the 12-year-old age group (Sandy *et al.*, 2001).

The research teams invited a total of 601 children to attend data collection days. 457 of these children were able to attend, of which 239 were five-year olds and 218 twelve-year olds. The standardised outcome records taken were:

- Dental study models
- Lateral cephalometric radiograph
- Anterior occlusal radiograph
- Clinical photographs
- Oral health assessment
- Patient satisfaction questionnaire
- Audio and video speech recordings
- Questionnaires to assess parent and patient satisfaction with clinical outcome

An overall low quality of outcome was observed. In the five-year-old cohort, 37% were scored as having a poor or very poor outcome using the 5-Year-Olds' Index, with 39% of 12-year-olds scoring as poor or very poor using the GOSLON yardstick. It is thought that children scoring as poor or very poor on the GOSLON yardstick are likely to require correction of a Class III skeletal relationship with orthognathic surgery once overall growth is complete. Forty two percent of secondary alveolar bone grafts performed were deemed to be unsuccessful, with 15% of the 12-year-old patients not having undergone the procedure

at all, despite all patients born with UCLP requiring it. This meant that the majority of patients in the UK born with UCLP either required a further operation or were left with a compromised dentition secondary to the failed or absent bone graft. Bone grafting at a later stage has been shown to have a poorer prognosis than grafts carried out at the ideal age of 9-11 years of age (Enemark *et al.*, 1988).

Oral health assessment revealed 5-year-old patients born with UCLP to have broadly similar caries rates to those of the general population, with 40% having caries requiring treatment, compared with 44% of 5-year-olds in the general population at that time (Pitts *et al.*, 1997). Thirty-nine percent of 5-year-olds and 10% of 12-year-olds had a persistent oral fistula causing symptoms including food lodging in the fistula, soft food escaping into the nose and regurgitation of fluids into the mouth.

Nineteen percent of the 12-year-old cohort of patients had speech that was either sufficiently different to provoke comment, was unintelligible to strangers, or impossible to understand (Sell *et al.*, 2001). Fifty one percent of the 5-year-old cohort had the same problems with speech, with only 20% assessed as having normal intelligibility. This contrasts with earlier findings estimating that half of patients with a repaired cleft palate develop normal speech without intervention (Spriestersbach *et al.*, 1973). The CSAG study reported that primary surgery for children with UCLP produces poor speech outcomes with the provision of speech therapy insufficient to meet need.

Despite the overall poor results, 12-year-old patients and their parents were almost all satisfied, or moderately satisfied, with the outcome of care. Only 9% were dissatisfied with the overall outcome of care.

Comparison between high volume operators (more than 10 surgeries per year) and low volume operators found improved outcomes for approximately one-third of key outcome variables for high volume operators. CSAG highlighted just how few cases most surgeons were treating with nearly 60% dealing with only one UCLP case per year. No significant differences in outcome were found between operators of different surgical specialities. When comparing the outcomes of the 12-year-old CSAG cohort to patients from the original 6 centre Eurocleft study, poorer outcomes were found in terms of dental arch relationships, midface retrusion and success of bone grafts (Bearn *et al.*, 2001).

2.2.2 CSAG recommendations

As a result of the CSAG study, the following recommendations were made (Bearn *et al.*, 2001):

1. Expertise and resources should be concentrated in 8 to 15 centres in the United Kingdom instead of the 57 operating at that time.
2. The range of expertise required in the team and the quality standards required should be clearly indicated by purchasers of care.
3. Units providing cleft care should ensure the full range of skills are available.
4. Clinicians should agree on a common nationwide database for all cleft patients.
5. Information on all cleft patients should be made available for comparative studies.

6. Training programs for all specialist cleft clinicians should be provided only in cleft centres where high-volume and high-quality clinical experience is available.
7. The surgical specialties involved must develop a common training pathway for the small number of trainees required to specialise in cleft care.
8. The Office of National Statistics should improve the recording of cleft births.

The UK government accepted all of these recommendations in February 1998, with the Department of Health acting to establish the cleft implementation group (CIG).

2.2.3 Changes following CSAG study

Implementation groups were set up across the UK, with meetings to seek public consultation on the proposals. To implement the recommended changes, cooperation from all members of the cleft teams was required, with reduction in both centres and personnel being necessary. As such, it is understandable that the implementation of changes proved to be a slow process. Only six centres had been chosen five years after the publication of the CSAG report (Murray, 2003). Centralisation has proved to be successful in improving outcomes in other specialities (Anderson *et al.*, 2011, Dikken *et al.*, 2012), with a large study of over 3 million patients demonstrating significantly lower mortality rates for patients undergoing a variety of different procedures in hospitals treating a high volume of patients, compared with those in lower volume hospitals (Reames *et al.*, 2014). Within cleft care, the two centres with the best outcomes in the Eurocleft study (Shaw *et al.*, 1992b), Denmark and Norway, were designated as national centres for the treatment of cleft lip and palate. It was suggested that a consistent surgical protocol used on a large number of patients was a

factor which contributed to improved growth outcomes relative to centres with the poorest outcomes in the study, both of which were regional cleft services based in the UK.

Prior to the CSAG study, 57 centres across the UK were providing cleft care. Following the CSAG recommendations, this number was reduced to 11 clinical networks by 2011 (Sandy *et al.*, 2012). A study performed soon after the implementation of a centralised model of care gave an early indication of how the new services were meeting the expectations of improved patient care (Hathorn *et al.*, 2006). This showed a substantial increase in the number of children in the two best outcome categories of the 5-Year-Olds' Index at 52%, compared with 29% in the CSAG study. This is close to the 55% in the two best centres from the original Eurocleft study. The Hathorn study also demonstrated a decrease in the number of children in the worst two outcome categories of the index at 22%, down from 37% in the CSAG study. It was noted that there was a need for better collection of records for audit purposes, as study models were only collected for 62% of the patient sample.

Data collected from six regional cleft centres for patients who had received an alveolar bone graft showed a successful radiographic outcome in 85% of cases, an improvement on the 58% reported in the CSAG study. However, different indices were used to measure the quality of the grafts (Revington *et al.*, 2010).

Following the CSAG recommendation that clinicians should agree on a national database, the Cleft Registry and Audit Network (CRANE) was set up in 2000 by the Department of Health to collect information about all children born with CLP in England, Wales and Northern Ireland. It is run by the Cleft Development Group, an independent body

representing patient representative groups, and funded by the NHS through the Specialist Service Commissioners, who have responsibility for the delivery of patient care. Demographic and treatment outcome related data is collected to facilitate national audit and research projects, and thoroughly report on the impact of care of patient outcomes.

2.2.4 Cleft Care UK (CCUK)

Fifteen years after the CSAG recommendations, the Cleft Care UK study from the University of Bristol examined the impact of centralisation on cleft services. In order to accurately compare results, the protocol was similar to that of the CSAG study, with extensions to include additional items. As not all 12-year-olds would have been cared for under a centralised service at the time of the study, this cross-sectional study was carried out on 5-year-olds born with UCLP only. The study recruited a total of 268 children born with UCLP.

Assessment of dental study models using the 5-Year-Olds' Index demonstrated a significant improvement in dentoalveolar outcomes in the CCUK group when compared with the CSAG group. 53% of models in the CCUK group were scored in categories 1 and 2, indicating good or excellent outcome, compared with 29.6% in the CSAG group. 19.2% of the CCUK models were scored in categories 4 and 5, indicating poor or very poor outcomes, compared with 36.3% in the CSAG group. When adjusted for age, the odds ratio for a better outcome was 2.29 (Al-Ghatam *et al.*, 2015).

The findings of the CCUK study indicate an improved outcome for facial appearance, with 36.2% rated as having a good or excellent appearance compared with 31.9% in the CSAG study. There was a decrease in children rated in the poor or very poor appearance category

from 27.6% in the CSAG group to 21.6% in the CCUK group. The odds ratio for an improved outcome for facial appearance was 1.43 (Al-Ghatam *et al.*, 2015).

There was strong evidence that speech outcomes were better in children in the CCUK group compared with those in the CSAG group, with improvement across most parameters (Sell *et al.*, 2015). There was an improvement in psychosocial outcomes, with 8% of parents reporting that their child's self-confidence had been adversely affected by their cleft, compared with 19% of parents in the CSAG study. Parental satisfaction with the care they had received from the cleft team improved from 93% to 98% (Waylen *et al.*, 2015).

The results in the areas of oral health were disappointing, with no improvement found in dmft rates and the prevalence of untreated caries remaining the same. Despite the recommendations in the CSAG report, only five of the 11 regional cleft units had been able secure funding for a consultant in paediatric dentistry, and of those only three were usually present at MDT clinics. Similarly, there were no improvements in audiology outcomes, with implementation of ENT and audiology into centralised multidisciplinary care slow and incomplete at the time of the CCUK study (Smallridge *et al.*, 2015).

The CCUK study demonstrated that a centralised multidisciplinary service improves outcomes, despite some areas still requiring improvement (Ness *et al.*, 2015). The surgeons in each of the 11 teams operate on a minimum of 35 cases per year (Fitzsimons *et al.*, 2012, Scott *et al.*, 2014, Scott *et al.*, 2015), a significant improvement compared with the large numbers of low volume operators across the previous 57 services providing cleft care. As a result of centralisation, conducting multicentre clinical research should be easier in the

future, further strengthening the evidence base to inform treatment decisions for children with CLP (Sandy *et al.*, 2012).

2.3 Clinical Governance

Clinical governance is an important part of clinical practice. It can be defined as '*a framework through which NHS organisations are accountable for continually improving the quality of their services and safeguarding high standards of care by creating an environment in which excellence in clinical care will flourish*' (Scally and Donaldson, 1998). It requires input at all levels of a healthcare organisation. Clinicians must be engaged, and service improvement generated through structures and processes that encompass clinical quality, service performance and financial control. There are seven key areas of activity which are used to ensure the NHS delivers high quality health care to service users (Picard and Wood, 2008):

- Clinical effectiveness and research
- Audit
- Risk management
- Education and training
- Patient and public involvement
- Using information and IT
- Staffing and staff management

The areas most relevant to this project are those of clinical effectiveness and clinical audit. Clinical effectiveness involves an evidence-based approach to clinical practice, research, developing new guidelines and protocols based on experience and evidence, and implementing guidelines and national standards to ensure optimal care. The aim of clinical

audit is to ensure clinical practice is continually monitored against set standards of best practice, highlighting areas where improvement is needed and adapting practice accordingly.

Outcome of clinical care is an important area to audit, measuring care against best practice standards established through research. In order to audit accurately, outcome measures must be used to quantify the relative success of an intervention.

2.3.1 Outcome measures

An important part of measuring any outcome of clinical care is selecting an appropriate outcome measure. The ideal properties of any such measure are, that it should be (Williams *et al.*, 2004):

- Valid (measures what it claims to measure)
- Reliable (measures the same thing in the same way on more than one occasion)
- Acceptable to the patient
- Non-invasive
- Precise (should be able to detect small changes in the condition)
- Clinically meaningful
- Easy to learn
- Quick to use
- Minimal equipment requirement
- Ability to use in the clinical setting
- Easy to record
- Evidence-based

The validity of the outcome measure is gauged by its sensitivity and specificity. Sensitivity is the ability of a test or outcome measure to correctly identify those with a condition, *i.e*. a test with 100% sensitivity correctly identifies all patients with the condition. Specificity refers to the ability of a test to correctly identify those patients without the condition (Lalkhen and McCluskey, 2008).

With regard to reliability, the outcome measure should demonstrate good inter-examiner reliability, resulting in the same score when used by two different examiners. It should also demonstrate good intra-examiner reliability, giving the same score when used by the same examiner on two different occasions.

2.4 Outcome measures used in CLP

Delivery of cleft care involves many specialties, each with at least one method of outcome assessment. As such, there are a large number of measures available, not all of which are necessarily perfectly reliable or valid, and so new ones are frequently being developed or modifications made to existing ones (Jones *et al.*, 2014). Different cleft phenotypes are not always assessed using the same outcome measures, as complexity and treatment varies. The majority of outcome measures developed to date are designed for use in assessing outcome of treatment in patients with UCLP. This is due to the fact that treatment involves both lip and palatal defects, and therefore covers outcomes across the full range of cleft care.

2.4.1 Speech

Palatal closure surgery is commonly carried out in the UK between the ages of 9 and 12 months, to try to facilitate normal anatomy and function. Delayed palatal closure is thought to minimise the negative effect of restricted maxillary growth, but speech outcomes are certainly significantly poorer (Rohrich *et al.*, 2000, Willadsen *et al.*, 2017, Willadsen *et al.*, 2018). CLP may cause various speech defects and therefore speech outcome measures must consider all aspects of speech.

The Great Ormond Street Speech Assessment (GOS.SP.ASS) was devised by Sell *et al.* (1994). It was later revised to address ambiguities in the original protocol (Sell *et al.*, 1999). It is a comprehensive tool for assessing speech in CLP patients by speech and language therapists, indicating the severity and location of speech errors. However, the tool is considered too time-consuming and too detailed for use as an audit tool (John *et al.*, 2006).

The cleft audit protocol for speech - augmented (CAPS-A) was developed for the purposes of audit by three cleft speech experts, who identified the key features required for existing assessment measures (John *et al.*, 2006). Elements of speech rated include intelligibility, nasality, errors in consonant production and other detailed elements. The outcomes are categorised on a traffic light system, along with need for surgery and speech and language therapy. The tool is found to be valid and reliable, and suitable for use in multicentre audit of speech outcomes for patients born with cleft palate. A comprehensive training programme for speech and language therapists was successfully developed in order to address the issue of standardising variables, allowing the tool to be used systematically and reliably (Sell *et al.*, 2009).

2.4.2 Nasolabial appearance

Improved facial appearance is an important goal of CLP treatment and facial asymmetry, as a result of CLP, causes significant emotional distress (Meyer-Marcotty and Stellzig-Eisenhauer, 2009, Meyer-Marcotty *et al.*, 2010). There are a large number of outcome measures (rating systems) available, which suggests a lack of general consensus and confidence in a reliable, valid, reproducible scoring system for assessing facial aesthetic outcomes in CLP (Sharma *et al.*, 2012).

The Asher-McDade indirect system assesses only the nasolabial area, using cropped photographs. Cropping was developed from a study which found that other facial features had a large influence on the perception attractiveness (Asher-McDade *et al.*, 1991). Therefore, cropped fronto-nasal and lateral views were used with the scale to assess nasal form, symmetry of the nose, shape of the vermillion border, nasal profile and upper lip. It was felt that a VAS was suited to relative, rather than absolute comparisons, and so a graduated 5-point ordinal scale was developed to reduce variation between examiners and improve interpretation. The system has been shown to be valid and has been used successfully in a number of multi-centre trials (Asher-McDade *et al.*, 1992, Mosmuller *et al.*, 2015). However, reliability is still questionable. Kuijpers-Jagtman *et al.* (2009) presented a set of 20 reference photographs, namely one for each of the five categories for each of the four views, in order to try to improve the reliability. Despite this reference set, the 2011 Americleft study found no significant increase in the reliability of the scoring system (Mercado *et al.*, 2011).

There are a several other facial aesthetic outcome measures, but none are as widely used as the Asher McDade system. The VLS (vermillion, lip, scar) method (Assuncao, 1992) is quick to use, but it has not been validated. Other measures include the cranio-facial proportion indices (Edler *et al.*, 2010), the aesthetic index (Johnson and Sandy, 2003) and the cleft lip evaluation profile (CLEP) (Ohannessian *et al.*, 2011). The Americleft project conducted a study to determine whether 3D images could be rated with comparable results and reliability, versus standard 2D clinical photographs. They found that 3D images were no more reliable than 2D images. The 2D images provided acceptable reliability and better accessibility for most cleft palate centres (Jones *et al.*, 2018). The Americleft project also proposed an expanded nasolabial appearance yardstick, for 5 to 7-year-old patients with UCLP, in an attempt to improve reliability (Mercado *et al.*, 2016). A Dutch centre developed the Cleft Aesthetic Rating Scale, which they report as being reliable and easy to use, although the results show no increase in reliability when compared with studies using the Asher McDade system (Mosmuller *et al.*, 2017).

2.4.3 Secondary alveolar bone grafting

Bone grafting is commonly performed as a secondary procedure for children with alveolar clefts involving the alveolus at around 8 to 10 years of age. Secondary alveolar bone grafting is a key procedure in comprehensive cleft care, with the primary objectives of stabilising the maxillary segments, improving vestibular soft tissue relationships, closing fistulae, facilitating tooth eruption, particularly the upper permanent canine, and to aid in obtaining nasal symmetry (Amanat and Langdon, 1991).

Measurement of success of bone grafting is generally assessed using standard radiographic images, either an upper occlusal or periapical. A variety of methods to assess success have been described, including measurement of interdental bone height (Bergland *et al.*, 1986), percentage of bony infill (Kindelan *et al.*, 1997) and the position of bone relative to adjacent root length (Witherow *et al.*, 2002). A comparison of these methods found reproducibility to be broadly similar across all three (Nightingale *et al.*, 2003). A study to compare examiner reliability of scoring radiographs following secondary bone grafting using a modified Kindelan Index and a 10-cm visual analogue scale (VAS) found better intra- and inter-examiner reliability using the VAS, which was more consistent regardless of level of clinical experience (Fowler *et al.*, 2018).

Measuring outcomes of surgery on three-dimensional bone using a two-dimensional image calls the validity into question. A systematic review of the clinical outcomes of secondary alveolar bone grafting using three-dimensional imaging concluded that the majority of findings in the literature were from observational studies with generally low methodological quality (De Mulder *et al.*, 2018).

A method of rating alveolar bone graft outcomes for CLP patients has recently been developed. The Standardised Way to Assess Grafts (SWAG) (Russell *et al.*, 2017) assesses both the quantity and location of bone within a grafted cleft side, providing further information which may help to determine the prognosis for regrafting, by identifying bony bridges and bony root coverage of adjacent teeth. Intra-rater reliability has been shown to be good to very good, with inter-rater reliability moderate to good. The SWAG scale was used in a study to assess secondary alveolar bone grafting outcomes in four cleft centres,

each with different protocols. It found levels of intra- and inter-rater reliability were good, and higher than those reported using other published methods (Russell *et al.*, 2016).

2.4.4 Patient satisfaction and quality of life

There is evidence that impaired facial growth and dental anomalies associated with CLP are linked with adverse psychosocial outcomes (Hunt *et al.*, 2005). Issues include difficulty with social relationships (Kramer *et al.*, 2009, Murray *et al.*, 2010), low self-confidence (Turner *et al.*, 1998) and an increased likelihood of being bullied (Shaw *et al.*, 1980).

Outcomes of CLP treatment are typically assessed objectively with observer or clinician-reported assessments (Semb *et al.*, 2005, Long *et al.*, 2011). However, the overall goal in CLP treatment is to improve a patient's health and quality of life (QOL). These outcomes are difficult to measure solely with observer and clinician-reported outcome measures. Patient Reported Outcome Measures (PROMs) can be used to evaluate a patient's perspective of their own health outcomes, quantifying quality of life and other significant outcome variables such as satisfaction, symptoms and function (Pusic *et al.*, 2011).

A systematic review conducted by Eckstein *et al.* (2011) found a lack of valid and reliable patient questionnaires for CLP. Five outcome measures validated for cleft populations were reported, but none were specifically created for clefts, and as such were of limited value. A subsequent systematic review (Klassen *et al.*, 2012) found no PROMs suitable for measuring the concerns of cleft affected individuals. They found areas of QOL research that were lacking, including satisfaction with appearance, cognitive function, family function, social function, social support, and school function.

CLEFT-Q (Tsangaris *et al.*, 2017) is a PROM that was developed following the review by Klassen *et al.* (2012) and is designed to measure outcomes that matter to affected individuals, including the domains of appearance, health related quality of life (HRQOL) and facial function. The validity of the questionnaires was determined using feedback from experts and interviews with patients. Interestingly the authors of CLEFT-Q reported the PROM had been field tested in an international study, but as yet no results have been published.

2.4.5 Dentoalveolar outcomes

Dentoalveolar outcome measures are an indirect measure of success of primary cleft surgery, as the surgery has an effect on maxillary growth and therefore the skeletal and dentoalveolar relationships. Outcome measures involve scoring study models against an index or set of reference models. There are a number of indices that have been developed over time and which are in use today. In general, they measure the anteroposterior relationships of the mandible and maxilla, with some also measuring vertical and transverse discrepancies.

2.4.5.1 Modified Huddart and Bodenham (MHB)

The Huddart and Bodenham system was developed for use in UCLP (Huddart and Bodenham, 1972), as it was felt existing indices were too subjective, with no assessment of validity or reliability. The index was developed on the primary dentition, with the position of each maxillary tooth scored relative to the opposing mandibular tooth, and the arch divided into right and left buccal and labial segments. This enabled the frequency and

severity of the dental crossbites to be scored, enabling evaluation of maxillary arch constriction. As the index is based directly on measurements it is entirely objective, with no requirement for calibration.

The index was later modified for use in BCLP (Heidbuchel and Kuijpers-Jagtman, 1997) and for the permanent dentition (Mossey *et al.*, 2003). It has been shown to be quick and easy to use, with high levels of reliability and sensitivity (Tothill and Mossey, 2007).

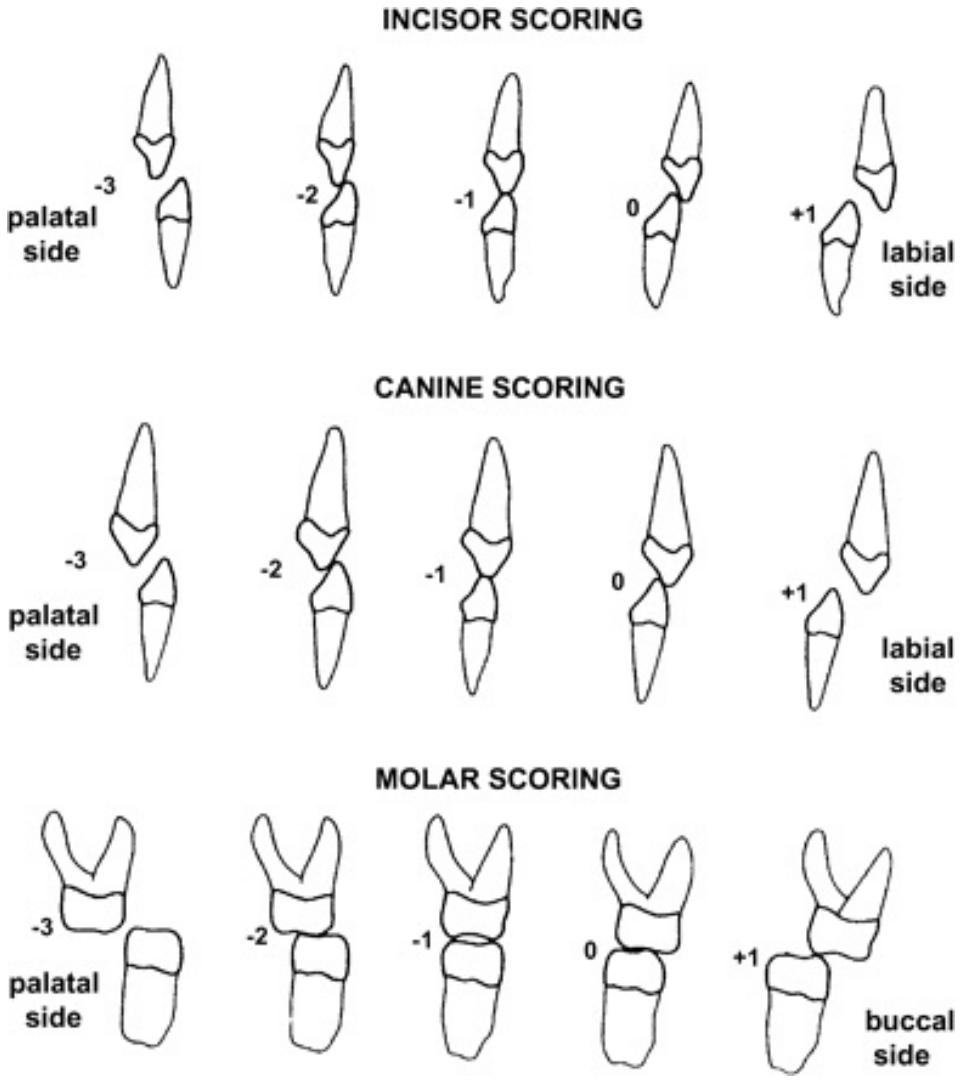


Figure 2: Modified Huddart and Bodenham scoring system (Dobbyn *et al.*, 2012)

Figure 2 illustrates the Modified Huddart and Bodenham scoring system used to assess the relationship of the maxillary and mandibular teeth. A retrospective study to describe and compare the use and reliability of the scoring system with the GOSLON and 5-Year-Olds' indices found a high level of intra- and inter-examiner reliability, with statistically significant correlation between scores from the three indices (Gray and Mossey, 2005). Dobbyn *et al.* (2012) carried out a study to correlate the range of MHB scores with those of the GOSLON and 5-Year-Old indices in order to create a scoring system that would allow a comparison with historical data. The study demonstrated excellent reliability and a high degree of correlation between MHB and GOSLON/ 5-Year-Olds' Index scores. The MHB demonstrated improved sensitivity compared with the other two indices, indicating that small changes in outcome can be measured more reliably. It equals or outperforms all other indices in the WHO criteria for an ideal index (Altalibi *et al.*, 2013). However, there are weaknesses of the MHB scoring system, namely that it weights both transverse and A-P discrepancies equally, does not include scoring for vertical discrepancies and ignores incisor inclination (Jones *et al.*, 2014)

2.4.5.2 GOSLON Yardstick

The Great Ormond Street, London and Oslo, Norway (GOSLON) Yardstick is the most widely used outcome measure for assessment of the results of primary surgery in 10-year-old patients born with UCLP (Mars *et al.*, 1987). It was developed as it was felt that the overall score of the Huddart and Bodenham system did not accurately represent the severity of the malocclusion, with the possibility of mild, generalised irregularity scoring more highly than a severe, but localised anomaly. The authors attempted to develop a simple and reliable

method of assessing the severity of the malocclusion regardless of assessor. The features of the malocclusion that were felt to be most important were:

1. A-P arch relationships
2. Vertical labial segment relationships
3. Transverse relationships

A-P arch relationships were considered to be the most important feature of the malocclusion, as this is the most difficult aspect to treat clinically. The index was applied to 30 sets of study models, from the archive of 12-year-old children at Great Ormond Street Hospital, chosen to represent the full range of results. Outcomes were categorised into five groups, ranging from excellent, with little or no orthodontic treatment required, to very poor, where orthognathic surgery will be required to obtain acceptable occlusal relationships. A set of reference models was produced to aid in categorisation of other study models by selecting one case from each of the five. This was then tested on 55 sets of models of children with UCLP from the Oslo Cleft Lip and Palate Clinic.

The Yardstick was shown to be extremely reliable and capable of discriminating the quality of results at different centres (Mars *et al.*, 1987). However, the authors themselves highlight the fact that the index is not able to finely discriminate between malocclusions. For the purposes of standardisation, clinicians must undergo training and calibration to use the GOSLON Yardstick, ensuring high levels of reliability between both operators and cleft units. There is a subjective element to assessment, and it is limited in that it can only be applied to children born with UCLP in the late mixed or early permanent dentition.

A systematic review concerning the predictive validity of the GOSLON Yardstick found that it is not capable of predicting growth patterns in patients born with UCLP (Buj-Acosta *et al.*, 2017). In one study, the GOSLON Yardstick failed to correctly predict growth in a third of cases (Jones *et al.*, 2016). A lack of longitudinal studies also means it is not possible to compare the predictive validity of the GOSLON Yardstick with other dentoalveolar outcomes in cleft affected individuals. However, it is the only index recommended for assessing outcomes of UCLP models at 10 years of age (Jones *et al.*, 2016).

2.4.5.3 5-Year-Olds' Index

Whilst the GOSLON Yardstick has proved to be a reliable and valid outcome measure of primary surgery, it is designed for use at 10 years of age, meaning a long period of time must elapse before the outcome of the primary surgery can be assessed. If it is required, secondary alveolar bone grafting usually occurs between the ages of 8 and 10, and some orthodontic treatment may also have occurred by this time. This means using the Yardstick to assess the results of primary surgery is subject to distortion from the subsequent procedures, resulting in an artificially improved score (Southall *et al.*, 2012).

The 5-Year-Olds' Index was developed in order to address these issues, enabling assessment of the outcome of primary surgery at the much earlier age of 5 years (Atack *et al.*, 1997a, Atack *et al.*, 1997b. This index, as shown in Table 2, kept the format and categories of the GOSLON Yardstick, but uses reference models of 5-year-old children born with UCLP and descriptions for the malocclusion in each category.

Groups	General features	Predicted long-term outcome
1	Positive overjet with average inclined or retroclined incisors No crossbites/open bites Good maxillary arch shape and palatal vault anatomy	Excellent
2	Positive overjet with average inclined or retroclined incisors Unilateral crossbite/crossbite tendency ± Open bite tendency around cleft site	Good
3	Edge-to-edge bite with average inclined or proclined incisors; or reverse overjet with retroclined incisors Unilateral crossbite ± Open bite tendency around cleft site"	Fair
4	Reverse overjet with average inclined or proclined incisors Unilateral crossbite ± bilateral crossbite tendency ± Open bite tendency around cleft site	Poor
5	Reverse overjet with proclined incisors Bilateral crossbite Poor maxillary arch form and palatal vault anatomy	Very poor

Table 2: 5-Year-Olds' Index scoring criteria (Atack *et al.*, 1997a, Atack *et al.*, 1997b

The reliability and validity of the 5-Year-Olds' index was assessed using the models of 60 children from both the Oslo Cleft centre and South West Cleft service (Atack *et al.*, 1997b). The study demonstrated excellent intra-examiner and good inter-examiner agreement, showing the index is both reproducible and reliable. When assessing the predictive validity of the 5-Year-Olds' Index, it was found to be on a par with the other indices, although once again the final occlusal outcome was only predicted in 50% of 5-year-old and 64% of 10-year old models (Jones *et al.*, 2016). It is therefore not possible to predict long-term growth from study models at the age of 5. There is an unconfirmed observation of systematic bias in using the index, which would also be the case with the GOSLON Yardstick. This is because

consultants experienced in cleft tend to score models more harshly than assessors who do not undertake cleft care. It is suggested that calibration prior to use of the index might improve consistency between examiners and reduce this systematic bias.

By using a tool that can reliably assess standards of care at the age of 5, it is easier to identify and audit the quality of primary surgery. This allows identification of techniques or surgical units providing the best or worst results at an early stage, giving an opportunity for any necessary alterations in care to be implemented at an earlier stage (Shaw *et al.*, 1992b).

2.4.5.4 Modified 5-Year-Olds' Index

As outcomes in cleft care have improved, it has become increasingly difficult to discriminate between the outcome categories of the 5-Year-Olds' Index, and therefore difficult to demonstrate continued improvements in outcomes. Research by Mittal (2018) aimed to develop and refine the original index to better discriminate between some of the outcome categories, using models selected from the CSAG and CCUK studies.

The verbal descriptors were used to develop the Modified 5-Year-Olds' Index (M5YO) as displayed in Table 3. Five categories were created from categories 1 to 3 of the original 5-Year-Olds' index in order to increase discrimination in these higher scoring categories. The lower scoring categories, 4 and 5, remained the same and became categories 6 and 7.

5-Year-Olds' Category	Modified Category	Features
1	1	Good +ve overjet Good +ve overbite Good archform Class II or I dentoalveolar
	2	Good +ve overjet Crossbite on C only Class II/2 or Class I incisors
2	3	+ve overjet Crossbite on some teeth in lesser segment (but some teeth not) Edge to Edge incisors with no crossbites
	4	Class III incisors Reducing overbite Nearly complete unilateral crossbite
3	5	Edge to Edge incisors Reduced/tenuous overbite Marked dentoalveolar compensation Unilateral crossbite
4	6	-ve overjet, incisors may be contacting Lower arch compensation Bilateral crossbite tendency Anterior openbite developing
5	7	Large rev overjet Bilateral crossbite Anterior openbite

Table 3: The Modified 5-Year-Olds' Index (Mittal *et al.*, 2018)

The validity of the Modified 5-Year-Olds' Index is comparable with that of the original 5-Year-Olds' Index (Mittal *et al.*, 2018). When compared with the original 5-Year-Olds' Index, a more even distribution of scores was demonstrated across the seven categories of the Modified 5-Year-Olds' Index. The modified index was found to be a reliable method of measuring the outcome of primary cleft surgery at the age of 5 years, and is able to discriminate more sensitively within the good outcome categories than the original index. Although the Modified 5-Year-Olds' Index was found to be reliable, testing was carried out using experienced examiners who also helped develop the index. This potential source of bias was noted by the author (Mittal *et al.*, 2018), who subsequently recommended further research be carried out to investigate the effect of experience when using the Modified 5-Year-Olds' Index.

2.5 Summary

Standards have improved across most outcome measures within cleft care in the UK following centralisation of services based on recommendations in the CSAG study (Sandy *et al.*, 1998). The 5-Year-Olds' Index is a reliable and valid tool to measure the dentoalveolar outcome of primary cleft surgery at the earliest age (Atack *et al.*, 1997a, Atack *et al.*, 1997b). The Modified 5-Year-Olds' Index was developed in order to better discriminate between the 'good' outcome categories of the original index. However, a calibration course and the use of reference models is a prerequisite for using the original 5-Year-Olds' Index. As such, for the Modified 5-Year-Olds' Index to be used to compare results between operators and cleft centres, it is important that a calibration course and reference models are established in order to fully develop the Modified 5-Year-Olds' Index.

3.0 AIMS AND OBJECTIVES

3.1 Research Aims

- To test the reliability of the Modified 5-Year-Olds' Index using examiners with a range of experience

3.2 Research Objectives

1. Assess the intra- and inter-examiner reliability of the Modified 5-Year-Olds' Index
2. Determine whether calibration is required for reliable use of the Modified 5-Year-Olds' Index
3. Determine whether level of orthodontic training has an effect on reliability when using Modified 5-Year-Olds' Index

4.0 MATERIALS AND METHODS

4.1 Sample selection

Study models were taken from the CCUK archives at the Bristol Dental School, University of Bristol. All of the records in this study were of non-syndromic children with a repaired unilateral cleft lip and palate aged between 5.3 and 6.5 years old. The patient selection criteria are described within the original CCUK study (Persson *et al.*, 2015); ethical approval for use of the models in the current research had been obtained as part of the original CCUK study application (REC reference number: 10/H0107/33, South West 5 REC). The proposal for the current project was approved by the Cleft Care UK Study Team.

4.2 Sample size calculation

The reliability of assessment tools such as the Modified 5-Year-Olds' Index must be established prior to their use in research or clinical applications. Reliability can be defined as the extent to which measurements can be replicated (Zapf *et al.*, 2016). For this study, intraclass correlation coefficients (ICCs) are used to measure reliability as they reflect both the degree of correlation and agreement between measurements.

As the aim of reliability studies is accurate estimation, the sample size calculation is based the estimated precision of the measure, *i.e.* the expected ICC, rather than power. The sample size calculation was performed based on confidence intervals as per the method proposed by Doros and Lew (2010), using PASS14 (Power Analysis and Sample Size System, NCSS Statistical Software. Kaysville, Utah, USA). Table 4 shows the sample size calculation, demonstrating that a sample of 50 study models will give a 95% confidence interval with a width of 0.209 when the expected ICC is 0.8.

Confidence Level	Number of subjects (n)	Observations per subject (k)	Width of confidence limits	Sample Intraclass Correlation (r)	Lower Confidence Limit (LCL)	Upper Confidence Limit (UCL)
0.950	50	2	0.209	0.800	0.673	0.881

Table 4: Figures used in the sample size determination for a two-way ICC (95% CI 0f 0.209 and an ICC of 0.8)

A sample of 50 models was therefore chosen from the 198 CCUK model set. The models were chosen to proportionally represent the spread of models in categories 1-7 of the Modified 5-Year-Olds' Index (Table 3) as scored by experts in consensus (Mittal *et al.*, 2018). The number of models per expert-scored category chosen are shown in Table 5.

Modified 5-Year-Olds' Category	Number of CCUK models selected for study
1	3
2	4
3	12
4	8
5	10
6	9
7	4

Table 5: Number of CCUK models chosen per Modified 5-Year-Olds' Index Category

The selected models were ordered at random using a random sequence generator *(https://www.random.org/sequences/)* and given a unique identifier (1-50). All data was entered onto an Excel Spreadsheet (Microsoft Office Corp, One Microsoft Way, Redmond, WA, USA).

4.3 Participant selection

Potential participants were identified and emailed a study information sheet (Appendix 1). They were identified to represent a broad range of experience, from second year specialty registrars still in training through to senior consultants with up to 16 years clinical experience in orthodontics, but none had significant experience working with individuals born with UCLP. A total of 15 participants agreed and were selected to take part in the study. The 15 participants were divided into three groups of equal experience by grade, each comprising one consultant, one post-CCST trainee and three orthodontic specialty registrars.

4.4 Groups

In order to determine whether use of reference models and/or calibration prior to using the Index would impact on reliability, the three groups were given differing amounts of information and resources prior to scoring the 50 models (Table 6).

Group	Resources	Assessors
Group 1	Information sheet only	Assessor 1 – consultant Assessor 2 – post-CCST trainee Assessor 3 – specialty trainee Assessor 4 – specialty trainee Assessor 5 – specialty trainee
Group 2	Information sheet Reference models	Assessor 6 – consultant Assessor 7 – post-CCST trainee Assessor 8 – specialty trainee Assessor 9 – specialty trainee Assessor 10 – specialty trainee
Group 3	Calibration course Information sheet Reference models	Assessor 11 – consultant Assessor 12 – post-CCST trainee Assessor 13 – specialty trainee Assessor 14 – specialty trainee Assessor 15 – specialty trainee

Table 6: Information/resources given to the three different groups in study

4.4.1 Information sheet

The information sheet was written for the purposes of the study with input from NEA, a consultant who devised the original 5-Year-Olds' Index and annually scores all models sent by cleft units from around the UK (Appendix 1). It is designed to clearly explain how to use

the Modified 5-Year-Olds' Index in order for participants to be able to use the Index without any other information.

A slightly extended version of this information sheet (Appendix 2) was produced for Groups 2 and 3, with an additional paragraph on how to use the reference models.

4.4.2 Reference models

A set of fourteen reference models for the Modified 5-Year-Olds' Index were retrieved from the archives at the University of Bristol. They were chosen by the expert examiners (NEA & SD) who were involved in development of the Modified 5-Year-Olds' Index (Mittal *et al.*, 2018). Two models were chosen for each of the seven categories to represent examples of expected features. As the reference models only represent examples of features that may be seen in each category, participants were instructed not to try and match these to the models in the study but to use them as a guide. Reference models were only provided to Groups 2 and 3.

4.4.3 Calibration course

The calibration course was developed with the help of NEA. It comprised a PowerPoint presentation (Microsoft Office Corp, One Microsoft Way, Redmond, WA, USA) using a selection of 20 models on which to practice scoring, taken from the CCUK archive but not otherwise used in this study (Appendix 3). The calibration course was delivered to Group 3 by NEA at Bristol Dental Hospital. Time was given for discussion, clarification, and debate on use of the Index, with advice provided by NEA on how best to categorise the models.

4.5 Model scoring

Each participant applied the Modified 5-Year-Olds' Index to the 50 study models and entered their scores onto a customised scoring sheet (Appendix 4). Each participant scored the models on two separate occasions, a minimum of four weeks apart, in order to minimise the effect of memory bias. After all participants had scored the models once, the models were renumbered (1 to 50), using a random sequence generator *(https://www.random.org/sequences/)*, again prior to rescoring and in order to minimise the effect of memory bias. Participants in Groups 2 and 3 had reference models laid out alongside models to be scored (Figure 3)

Figure 3: Photograph of the 50 study models, the 14 reference models and the information sheets

4.6 Statistical analysis

Statistical analysis was performed using Stata version 16 (StataCorp, College Station, Texas, USA) with a predetermined significance level of $\alpha = 0.05$. The data were analysed using Intraclass Correlation Coefficients two-way random effects model.

4.6.1 Intraclass Correlation Coefficients

ICCs are defined as 'True Variance' divided by 'Observed Variance'. In this case the 'True Variance' is the variability between the models, and the 'Observed Variance' the total variance minus true variance, plus other variance. Reliability values range between 0 and 1, with the strongest values closest to 1.

McGraw and Wong (1996) defined ten forms of ICC. These forms are based on the:

- Model – 1-way random effects, 2 way-random effects or 2-way fixed effects
- Type – single rater/measurement or the mean of raters/measurements (*k*)
- Definition of relationship considered to be important – consistency or absolute agreement

There are no standard values for acceptable reliability when using ICCs. A low ICC in this case could represent several different variables – lack of rater agreement, lack of model variability, the small number of models and the small number of raters.

4.6.2 Two-way random effects model

The aim of this study was to test the reliability of the Modified 5-Year-Olds' Index using examiners with a range of experience in the field of orthodontics. The examiners in this

study were selected from a wider population of orthodontists and orthodontic trainees, so the examiner is considered one random factor, the other being variability between the study models. The two-way random effects model was chosen for calculating ICCs in this study as this form of ICC means the reliability results can be generalised to any raters possessing the same characteristics as the raters in this study. For comparison, the one-way random effects model is used when each subject is rated by a different set of raters randomly chosen from a wider population of potential raters. The two-way mixed effects model is used if the selected raters are the only raters of interest i.e. the results cannot be generalised to other raters even if the rater characteristics are similar (Koo and Li, 2016).

ICCs calculated from a two-way random effects model were used to assess:

- The intra-rater reliability of individual examiner's rankings of models in each category.
- The inter-rater reliability in comparing an examiner's ranking to the expert's consensus rank (gold standard) in the original Modified 5-Year-Olds' Index study (Mittal *et al.*, 2018).

5.0 RESULTS

5.1 Intra-rater reliability

50 models from the CCUK cohort were scored by 15 assessors across the three groups. Group 1 were provided with an information sheet on using the Index; Group 2 had the information sheet and a set of fourteen reference models; Group 3 attended a calibration course prior to scoring and had use of the information sheet and reference models. Comparison of agreement between scoring sessions was assessed using ICCs based on a two-way random effects model (Table 7).

	Consultant (Assessors 1, 6 and 11)	**Post-CCST trainee (Assessors 2, 7 and 12)**	**Specialty Trainee (Assessors 3, 8 and 13)**	**Specialty Trainee (Assessors 4, 9 and 14)**	**Specialty Trainee (Assessors 5, 10 and 15)**
Group 1	0.93 (0.88, 0.96)	0.89 (0.81, 0.94)	0.78 (0.64, 0.87)	0.68 (0.50, 0.81)	0.72 (0.55, 0.83)
Group 2	0.87 (0.80, 0.93)	0.87 (0.77, 0.93)	0.80 (0.67, 0.88)	0.88 (0.79, 0.92)	0.93 (0.88, 0.96)
Group 3	0.83 (0.70, 0.90)	0.91 (0.83, 0.95)	0.91 (0.85, 0.95)	0.80 (0.67, 0.88)	0.83 (0.70, 0.90)

Table 7: Intra-rater ICC values and 95% CIs for each assessor in each of the three groups

It can be seen that Group 1 demonstrates the overall lowest levels of intra-rater reliability, with the lowest three scores seen being those of the specialty trainees of the group. Assessor 4 in Group 1 had the lowest overall score, with an ICC of 0.68 and the widest 95% confidence interval of the assessors at 0.50 to 0.81. The consultant and post-CCST trainee

of Group 1 had intra-rater reliability scores comparable with those of Groups 2 and 3. The two assessors with the highest levels intra-rater reliability were the consultant of Group 1 and specialty trainee 3 in Group 2, with an ICC and 95% CI of 0.93 (0.88, 0.96).

The number of models scored in each category for each examiner at both time points are shown by cross-tabulation in Tables 8-22. The cross-tabulated scores demonstrate the agreement of all assessors between Sessions 1 and 2. The numbers in the blue boxes demonstrate the number of models with perfect agreement in each category. The further away from the blue boxes the numbers are, the greater the disagreement and by how many Index categories.

The vast majority of models without perfect agreement across sessions were scored only one category apart. The majority of these disagreements occur between categories 3 and 6.

5.1.1 Group 1 cross-tabulated scores

Assessor 1 – Session 1	Assessor 1 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2							2
	2		5						5
	3			7	7				14
	4				5	4			9
	5					7	1		8
	6					1	7	1	9
	7						2	1	3
	Total	2	5	7	12	12	10	2	50

Table 8: Cross tabulation for scores between sessions for Assessor 1 (Consultant) in Group 1 using the Modified 5-Year-Olds' Index

Assessor 2 – Session 1	Assessor 2 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	1		1					2
	2		2	1					3
	3		2	9	3	1			15
	4				3	1			4
	5				4	8	1		13
	6					2	7	1	10
	7						2	1	3
	Total	1	4	11	10	12	10	2	50

Table 9: Cross tabulation for scores between sessions for Assessor 2 (Post-CCST trainee) in Group 1 using the Modified 5-Year-Olds' Index

Assessor 3 – Session 1	Assessor 3 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2							2
	2		1	1					2
	3		2	8	3		1		14
	4			2	1	2			5
	5			1	1	4	2		8
	6		1	1		2	9		13
	7					1	3	2	6
	Total	2	4	13	5	9	15	2	50

Table 10: Cross tabulation for scores between sessions for Assessor 3 (Specialty Trainee) in Group 1 using the Modified 5-Year-Olds' Index

Assessor 4 – Session 1	Assessor 4 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	1							1
	2		1	2					3
	3				1				1
	4		1	2	5	5	1		14
	5		1	3	8	6	1		19
	6				1	1	5		7
	7					1	1	3	5
	Total	1	3	7	15	13	8	3	50

Table 11: Cross tabulation for scores between sessions for Assessor 4 (Specialty Trainee) in Group 1 using the Modified 5-Year-Olds' Index

Assessor 5 – Session 1	Assessor 5 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	1							1
	2			3	1				4
	3		2	5	3	2		1	13
	4		2	2	1		2		7
	5		1		2	4	4		11
	6					3	6	1	10
	7							4	4
	Total	1	5	10	7	9	12	6	50

Table 12: Cross tabulation for scores between sessions for Assessor 5 (Specialty Trainee) in Group 1 using the Modified 5-Year-Olds' Index

The speciality trainees of Group 1 show the overall highest number of disagreements across sessions, with some models being scored particularly far apart in terms of categories (Tables 10-12). One of the specialty trainees, Assessor 5, scored one of the models a category 3 in one session and category 7 in the other (Table 12); another specialty trainee, Assessor 3, scored one of the models a category 2 and a category 6 (Table 10).

5.1.2 Group 2 cross-tabulated scores

Assessor 6 – Session 1	Assessor 6 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2							2
	2	1	3	2	1				7
	3		1	1	4	1			7
	4		1	1	7	1			10
	5				3	3	2		8
	6					5	6	2	13
	7							3	3
	Total	3	5	4	15	10	8	5	50

Table 13: Cross tabulation for scores between sessions for Assessor 6 (Consultant) in Group 2 using the Modified 5-Year-Olds' Index

Assessor 7 – Session 1	Assessor 7 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		4						4
	3		4	5	2	1			12
	4			2	4				6
	5			2	5	5	2		14
	6					4	5		9
	7						1	1	2
	Total	3	8	9	11	10	8	1	50

Table 14: Cross tabulation for scores between sessions for Assessor 7 (Post-CCST trainee) in Group 2 using the Modified 5-Year-Olds' Index

Assessor 8 – Session 1	Assessor 8 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2							2
	2	1	3						4
	3		1	9	2				12
	4		1	2	2	2	1		8
	5			1	1	7	2		11
	6		1		1	2	5		9
	7					1	2	1	4
	Total	3	6	12	6	12	10	1	50

Table 15: Cross tabulation for scores between sessions for Assessor 8 (Specialty Trainee) in Group 2 using the Modified 5-Year-Olds' Index

Assessor 9 – Session 1	Assessor 9 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		3						3
	3		1	13	1	1			16
	4			4	3	1			8
	5				3	2	1		6
	6				1	2	8	1	12
	7				1			1	2
	Total	3	4	17	9	6	9	2	50

Table 16: Cross tabulation for scores between sessions for Assessor 9 (Specialty Trainee) in Group 2 using the Modified 5-Year-Olds' Index

Assessor 10 – Session 1	Assessor 10 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		2	2	1				5
	3		1	8	3				12
	4			3	3	2			8
	5				1	5	1		7
	6					2	9		11
	7							4	4
	Total	3	3	13	8	9	10	4	50

Table 17: Cross tabulation for scores between sessions for Assessor 10 (Specialty Trainee) in Group 2 using the Modified 5-Year-Olds' Index

In Group 2, the majority of models were scored in either the same category or one category apart (Tables 12-17). All assessors scored a small number of models two categories apart between sessions. Assessor 9 scored one model three categories apart (Table 16); Assessor 8 scored one model four categories apart (Table 15).

5.1.3 Group 3 cross-tabulated scores

Assessor 11 – Session 1	Assessor 11 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2	1	2	3	1				7
	3			5	10		1		16
	4			1	2	4			7
	5				1	5	1		7
	6				1	2	6		9
	7							1	1
	Total	4	2	9	15	11	8	1	50

Table 18: Cross tabulation for scores between sessions for Assessor 11 (Consultant) in Group 3 using the Modified 5-Year-Olds' Index

Assessor 12 – Session 1	Assessor 12 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3	1						4
	2		2						2
	3			17	1				18
	4				1				1
	5			2	2	4			8
	6				2	5	8		15
	7							2	2
	Total	3	3	19	6	9	8	2	50

Table 19: Cross tabulation for scores between sessions for Assessor 12 (Post-CCST trainee) in Group 3 using the Modified 5-Year-Olds' Index

Assessor 13 – Session 1	Assessor 13 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2							2
	2	1		1	1				3
	3		1	5	2				8
	4			1	7	4			12
	5					13	2		15
	6					1	6		7
	7						1	2	3
	Total	3	1	7	10	18	9	2	50

Table 20: Cross tabulation for scores between sessions for Assessor 13 (Specialty Trainee) in Group 3 using the Modified 5-Year-Olds' Index

Assessor 14 – Session 1	Assessor 14 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2		1					3
	2		2	2	1				5
	3		1	10	2	2			15
	4			4	3	1			8
	5				6	4	1		11
	6				1	2	4		7
	7							1	1
	Total	2	3	17	13	9	5	1	50

Table 21: Cross tabulation for scores between sessions for Assessor 14 (Specialty Trainee) in Group 3 using the Modified 5-Year-Olds' Index

Assessor 15 – Session 1	Assessor 15 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	0	2						2
	2	1	1	1					3
	3		2	4	1				7
	4			3	4	2			9
	5			2	3	4	3		12
	6				1	8	5		14
	7						1	2	3
	Total	1	5	10	9	14	9	2	50

Table 22: Cross tabulation for scores between sessions for Assessor 15 (Specialty Trainee) in Group 3 using the Modified 5-Year-Olds' Index

The cross-tabulated scores of Group 3 (Tables 18-22), calibrated with access to information sheet and reference models, show similar levels of agreement in scoring models to Group 2. The majority of models were scored with agreement between sessions or no more than one category apart. All assessors scored a small number of models two categories apart across the two scoring sessions. Only one individual, Assessor 11, scored a model three categories apart between the sessions (Table 18).

5.2 Inter-examiner reliability

Comparison of agreement between examiners and the gold standard was assessed using ICCs based on a two-way random effects model. The gold standard is the consensus score for each model as agreed by expert examiners (NEA, SAD) during development of the Index. ICCs and 95% CIs were calculated for the two scoring sessions which took place at least four weeks apart (Tables 23 + 39).

5.2.1 Session 1 inter-examiner ICCs and 95% CIs

	Consultant (Assessors 1, 6 and 11)	Post-CCST trainee (Assessors 2, 7 and 12)	Specialty Trainee (Assessors 3, 8 and 13)	Specialty Trainee (Assessors 4, 9 and 14)	Specialty Trainee (Assessors 5, 10 and 15)
Group 1	0.94 (0.90, 0.96)	0.92 (0.86, 0.95)	0.77 (0.62, 0.86)	0.69 (0.45, 0.82)	0.86 (0.77, 0.92)
Group 2	0.93 (0.88, 0.96)	0.92 (0.86, 0.95)	0.92 (0.87, 0.96)	0.91 (0.85, 0.95)	0.90 (0.83, 0.94)
Group 3	0.90 (0.68, 0.96)	0.91 (0.84, 0.95)	0.88 (0.80, 0.93)	0.88 (0.76, 0.94)	0.84 (0.70, 0.91)

Table 23: ICC values and 95% CIs between examiners and the gold standard score for the Modified 5-Year-Olds' Index for Session One

In Group 1, the information sheet only group, the ICCs ranged from 0.69 to 0.94, displaying the least consistency between assessors compared with the other two groups (Table 23). The widest 95% confidence interval was seen for one of the specialty trainees (Assessor 4) in this group (0.45, 0.82), demonstrating a low level of reliability. However, the consultant (Assessor 1) in Group 1 had the highest level of reliability of all assessors across all groups, with the highest ICC and the narrowest 95% confidence interval. The post-CCST trainee (Assessor 2) also demonstrated high levels of reliability with a high ICC and relatively narrow 95% confidence interval.

The ICCs of Group 2, information and reference model group, ranged from 0.90 to 0.93, displaying the overall highest levels of reliability across the three groups with relatively narrow 95% confidence intervals. The ICCs of Group 3 were slightly lower than those in

Group 2, ranging from 0.84 to 0.91, but the 95% confidence intervals were also significantly wider.

Assessor scores are cross-tabulated against the gold standard expert consensus scores for both session one (Tables 24-38) and session two (Appendix 5, Tables 40-54). The boxes highlighted in blue indicate the expected category as per the gold standard score. The further away from the blue box the numbers are, the greater the disagreement and by how many index categories.

The cross-tabulated assessor scores against the gold standard expert consensus scores show a similar pattern to the intra-rater cross-tabulated scores. The specialty trainees of Group 1 (Assessors 3-5) show the highest number of disagreements with the gold standard score. They had the highest number of scores that deviated from the gold standard by more than one category; they were the only assessors to score a model four categories apart from the gold standard, and this occurred on three occasions (Tables 26 and 27).

There were good levels of agreement with the gold standard, with over half of all models scored in agreement across all groups. The cross-tabulated scores demonstrate that levels of agreement are highest overall in Group 2, with Group 3 close behind (Tables 29-38). Levels of agreement overall for Group 1 were the lowest, with the agreement levels of the specialty trainees considerably lower than those of the consultant and post-CCST trainee. The majority of models that were not scored in agreement with the gold standard scored only one category either side, with the majority of the disagreements occurring between categories 3 and 6.

5.2.2 Group 1 Session 1 cross-tabulated scores

Gold standard consensus score	Assessor 1 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2	1						3
	2		3	1					4
	3		1	9	2				12
	4			4	4				8
	5				3	6	1		10
	6					2	7		9
	7						1	3	4
	Total	2	5	14	9	8	9	3	50

Table 24: Cross tabulation of Assessor 1 (Consultant) scores for session 1 and gold standard expert consensus score

Gold standard consensus score	Assessor 2 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2	1						3
	2		2	2					4
	3			11	1				12
	4			2	2	3			7
	5				1	9		1	11
	6					1	8		9
	7						2	2	4
	Total	2	3	15	4	13	10	3	50

Table 25: Cross tabulation of Assessor 2 (post-CCST trainee) scores for session 1 and gold standard expert consensus score

Gold standard consensus score	Assessor 3 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2	1						3
	2		1	2			1		4
	3			8	2	1		1	12
	4			3	3	1		1	8
	5			1		4	5		10
	6					2	7		9
	7							4	4
	Total	2	2	14	5	8	13	6	50

Table 26: Cross tabulation of Assessor 3 (specialty trainee) scores for session 1 and gold standard expert consensus score

Gold standard consensus score	Assessor 4 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	1	2						3
	2		1		2	1			4
	3			1	6	4		1	12
	4				3	5			8
	5				1	8	1		10
	6				2	1	6		9
	7							4	4
	Total	1	3	1	14	19	7	5	50

Table 27: Cross tabulation of Assessor 4 (specialty trainee) scores for session 1 and gold standard expert consensus score

Gold standard consensus score	Assessor 5 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	1	1	1					3
	2		2	1	1				4
	3		1	8	2	1			12
	4			3	2	3			8
	5				1	4	5		10
	6				1	3	5		9
	7							4	4
	Total	1	4	13	7	11	10	4	50

Table 28: Cross tabulation of Assessor 5 (specialty trainee) scores for session 1 and gold standard expert consensus score

5.2.3 Group 2 Session 1 cross-tabulated scores

Gold standard consensus score	Assessor 6 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2	1						3
	2		3	1					4
	3		3	4	5				12
	4			2	5	1			8
	5					6	4		10
	6					1	8		9
	7						1	3	4
	Total	2	7	7	10	8	13	3	50

Table 29: Cross tabulation of Assessor 6 (Consultant) scores for session 1 and gold standard expert consensus score

Gold standard consensus score	Assessor 7 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		3	1					4
	3		1	8	2	1			12
	4			2	4	2			8
	5			1		8	1		10
	6					3	6		9
	7						2	2	4
	Total	3	4	12	6	14	9	2	50

Table 30: Cross tabulation of Assessor 7 (post-CCST trainee) scores for session 1 and gold standard expert consensus score

Gold standard consensus score	Assessor 8 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2	1						3
	2		3	1					4
	3			10	1		1		12
	4			1	6	1			8
	5				1	7	2		10
	6					3	6		9
	7							4	4
	Total	2	4	12	8	11	9	4	50

Table 31: Cross tabulation of Assessor 8 (specialty trainee) scores for session 1 and gold standard expert consensus score

Gold standard consensus score	Assessor 9 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		3	1					4
	3			10	1	1			12
	4			5	3				8
	5				3	4	3		10
	6				1	1	7		9
	7						2	2	4
	Total	3	3	16	8	6	12	2	50

Table 32: Cross tabulation of Assessor 9 (specialty trainee) scores for session 1 and gold standard expert consensus score

Gold standard consensus score	Assessor 10 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		3	1					4
	3		2	7	2		1		12
	4			3	4	1			8
	5			1	2	4	3		10
	6					2	7		9
	7							4	4
	Total	3	5	12	8	7	11	4	50

Table 33: Cross tabulation of Assessor 10 (specialty trainee) scores for session 1 and gold standard expert consensus score

5.2.4 Group 3 Session 1 cross-tabulated scores

Gold standard consensus score	Assessor 11 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2	1						3
	2		4						4
	3	1	2	9					12
	4			7	1				8
	5				5	5			10
	6				1	2	6		9
	7						3	1	4
	Total	3	7	16	7	7	9	1	50

Table 34: Cross tabulation of Assessor 11 (Consultant) scores for session 1 and gold standard expert consensus score

Gold standard consensus score	Assessor 12 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2	1	2	1					4
	3			11		1			12
	4			6			2		8
	5				1	6	3		10
	6					1	8		9
	7						2	2	4
	Total	4	2	18	1	8	15	2	50

Table 35: Cross tabulation of Assessor 12 (post-CCST trainee) scores for session 1 and gold standard expert consensus score

Gold standard consensus score	Assessor 13 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2	1						3
	2		1	2	1				4
	3		1	5	6				12
	4			1	3	4			8
	5				1	8	1		10
	6				1	3	5		9
	7						1	3	4
	Total	2	3	8	12	15	7	3	50

Table 36: Cross tabulation of Assessor 13 (specialty trainee) scores for session 1 and gold standard expert consensus score

Gold standard consensus score	Assessor 14 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2	1						3
	2	1	2	1					4
	3		2	10					12
	4			3	4	1			8
	5				4	5	1		10
	6			1		5	3		9
	7						3	1	4
	Total	3	5	15	8	11	7	1	50

Table 37: Cross tabulation of Assessor 14 (specialty trainee) scores for session 1 and gold standard expert consensus score

Gold standard consensus score	Assessor 15 – Session 1								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2	1						3
	2		2	1		1			4
	3			5	5	1	1		12
	4			1	3	4			8
	5				1	5	4		10
	6					1	8		9
	7						1	3	4
	Total	2	3	7	9	12	14	3	50

Table 38: Cross tabulation of Assessor 15 (specialty trainee) scores for session 1 and gold standard expert consensus score

5.2.5 Session 2 inter-examiner ICCs and 95% CIs

	Consultant	**Senior**	**Junior 1**	**Junior 2**	**Junior 3**
Group 1	0.93 (0.88, 0.96)	0.90 (0.82, 0.94)	0.81 (0.69, 0.89)	0.75 (0.60, 0.85)	0.83 (0.71, 0.90)
Group 2	0.92 (0.87, 0.96)	0.90 (0.79, 0.95)	0.88 (0.79, 0.93)	0.87 (0.76, 0.93)	0.89 (0.81, 0.94)
Group 3	0.82 (0.70, 0.89)	0.90 (0.81, 0.95)	0.90 (0.81, 0.94)	0.80 (0.65, 0.89)	0.89 (0.82, 0.94)

Table 39: ICC values and 95% CIs between examiners and the gold standard score for the Modified 5-Year-Olds' Index for Session Two

The pattern of results for the second scoring session were extremely similar to the first. The consultant in Group 1 again had the highest ICC at 0.93 with the narrowest 95% confidence

interval, demonstrating a high level of reliability (Table 39). The post-CCST trainee had an ICC of 0.90 and a relatively narrow 95% confidence interval, again demonstrating high levels of reliability. Group 1 had lowest overall range of ICCs, from 0.75 to 0.93, with the overall widest confidence intervals, again as a result of the specialty trainees having the lowest ICCs and widest 95% confidence intervals.

Group 2, the information sheet and reference model group, had the highest overall ICCs across the three groups with a range of 0.87 to 0.92, though the 95% confidence intervals were wider than those in session one. The overall ICCs were again slightly lower for Group 3 compared with Group 2, ranging from 0.82 to 0.90.

5.3 Reliability by group

The group with the overall highest levels of intra-rater reliability was Group 2. The assessors in Group 2 had the information sheet on use of the Index and access to the set of reference models for use during model scoring, with ICCs ranging from 0.80 (0.67, 0.88) to 0.93 (0.88, 0.96) (Table 7). Group 3 had intra-rater reliability ICCs only marginally lower than Group 2, ranging from 0.80 (0.67, 0.88) to 0.91 (0.85, 0.95), even though they had been calibrated and it might have been expected they would have had the highest level of reliability. Intra-rater reliability ICCs for Group 1 ranged from 0.68 (0.50, 0.81) to 0.93 (0.88, 0.96), demonstrating the least reliability of the three groups. This was perhaps to be expected, with assessors scoring models with no prior experience of the index and only an information sheet to refer to for guidance.

When comparing scores between assessors and the gold standard score for the Modified 5-Year-Olds' Index, Group 2 demonstrated the highest level of inter-rater reliability for both the first and second scoring sessions, with a washout period of at least 4 weeks in between the two (Tables 23 and 39). The inter-rater ICC scores for the two sessions of Group 3 were both slightly lower than those of Group 2. The overall range of ICCs for Group 1 were considerably lower than those of Groups 2 and 3 and ICC scores at the lower end in this group exhibited particularly low reliability.

Whilst there are differences in intra- and inter-rater reliability ICC scores between groups, in fact the only significant differences were found at the specialty trainee level.

5.4 Reliability by level of training

The level of training of the assessors (i.e. Consultant, post-CCST trainee or specialty trainee) appears to influence reliability scores in Group 1. In this group, the consultant and post-CCST trainee both had high intra-rater reliability scores, with scores very similar to the consultants and post-CCST trainees in the other groups (Table 7). The range of intra-rater ICC scores for the specialty trainees in Group 1 was considerably lower compared with the scores of those at the same level of training in the other groups. The intra-rater ICC scores of the specialty trainees in Groups 2 and 3 were very similar and are in fact within the same range as the consultants and post-CCST trainees of all groups.

The same pattern was seen as per the intra-rater reliability ICCs when comparing assessor scores with the gold standard, with the range of inter-rater ICCs for specialty trainees in

Group 1 lower than the scores for all assessors in other groups in both scoring sessions (Tables 23 and 39).

6.0 DISCUSSION

Outcomes in cleft care improved following the implementation of a centralised model of care as recommended by the CSAG study (Bearn *et al.*, 2001). The dentoalveolar outcomes assessed using the 5-Year-Olds' Index (Al-Ghatam *et al.*, 2015, Ness *et al.*, 2015) demonstrated significant improvements in the centralised service. These changes resulted in a greater number of study models scored in the higher categories of the 5-Year-Olds' Index, and there was therefore a need to be able to discriminate between outcomes with a greater degree of sensitivity. As a result the Modified-5-Year-Olds' Index (Mittal *et al.*, 2018) was developed with the expansion from five to seven categories; categories 1-3 of the original index were expanded to five categories to increase discrimination in the higher scoring categories, with the lower scoring categories 4 and 5 becoming 6 and 7.

The Modified 5-Year-Olds' Index is a reliable method of measuring the outcome of primary cleft surgery at the age of 5 years and is able to discriminate more sensitively within the good outcome categories than the original Index (Mittal *et al.*, 2018). However, reliability testing of the Modified 5-Year-Olds' Index was carried out by expert assessors with a number of years of experience in the field of cleft orthodontics. They also played a significant role in development of the index (Mittal *et al.*, 2018). The original 5-Year-Olds' Index required calibration to be used reliably (Atack *et al.*, 1997a). The aim of this study was to determine whether calibration is required for reliable use of the Modified 5-Year-Olds' Index, and whether the level of training of the assessor has any effect on reliability.

6.1 The assessors

Fifteen people working in secondary care orthodontic departments around the South West of England volunteered to take part in the study. Those selected to participate were:

- Three consultants
- Three post-CCST trainees
- Nine specialty trainees

It was important to select assessors with a broad range of experience in this study. Testing of the Modified 5-Year-Olds' Index to date demonstrated excellent intra- and inter-rater reliability (Mittal *et al.*, 2018). However, this could be due to fact there were only two raters, both of whom had extensive experience in using the original 5-Year-Olds' Index and were heavily involved in development of the Modified 5-Year-Olds' Index. Selecting raters for this study with experience ranging from consultants, with up to 16 years in clinical orthodontics, to specialty trainees in their second year of training enabled comparison in levels of reliability between experienced and inexperienced examiners.

6.2 Ease of use of indices

There are a variety of indices used to measure dentoalveolar outcomes in cleft care as described in section 2.4.5. For a new index to be widely adopted, it must be easy to use, otherwise clinicians are likely to stick to alternatives they are already familiar with. Ease of use of an index can be judged on a number of factors including the time it takes to score models, user friendliness, equipment required such as reference models, and training required prior to use of the index (Jones *et al.*, 2016).

Whilst not formally timed, none of the assessors took more than one hour to score all 50 models, including the time to read the information sheet and become familiar with a new index prior to commencing model scoring. Some of the more experienced assessors were able to score all models in around 20 minutes. A previous study comparing indices used to measure the dentoalveolar outcome of primary cleft surgery (Jones *et al.*, 2016) found that the quickest index to use in measuring the study models of the 5-year-old cohort was simple overjet measurement as described by Morris *et al.* (1994). However, this index is not comparable in its ability to measure the full range dentoalveolar outcomes, as it measures only overjet and does not consider the occlusion in three dimensions, calling its validity into question. The GOSLON Yardstick (Mars *et al.*, 1987) was next quickest to use, and this is the most widely used dentoalveolar outcome measure of primary surgery (Hathaway *et al.*, 2011). The original 5-Year-Olds' Index was not far behind the GOSLON Yardstick in terms of speed of use, and the expert assessors testing this new Modified 5-Year-Olds' Index felt that it did not take much longer to use the modified index in comparison to the original index (Mittal *et al.*, 2018). The Modified Huddart Bodenham Index (Huddart and Bodenham, 1972) and the EUROCRAN (Fudalej *et al.*, 2011) indices took the longest to complete due to their increased complexity.

The need for training or calibration prior to use of an index has an impact on its ease of use (Dobbyn *et al.*, 2012). The need for calibration means potential assessors need to take the time to find and attend a calibration course prior being able to use the index reliably. This may reduce the number of clinicians using the index and therefore have a negative impact on its wider uptake. The 5-Year-Olds' Index requires calibration for the purpose of standardisation, ensuring high levels of reliability in scoring between operators and units

and reducing systematic bias (Atack *et al.*, 1997b). Interestingly the results of the present study show that calibration does not improve reliability in use of the Modified 5-Year-Olds' Index. When comparing Group 2 (information sheet and reference models) and Group 3 (information sheet, reference models and calibration course), Group 2 had intra-rater ICCs ranging from 0.80 to 0.93 compared to Group 3 ICCs ranging from 0.80 to 0.91 (Table 7). When comparing assessor scores to the gold standard scores for the ICCs, for Group 2 they ranged from 0.87 – 0.93, and for Group 3 from 0.80 – 0.91 (Tables 23 and 39). The above scores demonstrate high levels of agreement across all assessors, but slightly lower numbers for those that had been calibrated prior to scoring the models.

The 5-Year-Olds' Index and the GOSLON Yardstick both require a set of reference models for use (Dobbyn *et al.*, 2012, Jones *et al.*, 2016). The results of the current study show that reference models are required for the Modified 5-Year-Olds' Index to be used reliably by assessors of all levels in orthodontics. This might be considered as a barrier for use of an index since the reference models must be obtained and stored long-term. In the medium- to long-term this is a problem that may be overcome with a move towards using three-dimensional digital study models in orthodontics. This would provide instant access to models without the need for large amounts of storage space, the ability to transfer them anywhere worldwide quickly and securely, eliminating the risk of breakage or loss (Fleming *et al.*, 2011).

Previous studies have compared scoring using 3D models versus plaster models with the 5-Year-Olds' Index and the GOSLON Yardstick (Dogan *et al.*, 2012, Chawla *et al.*, 2013, Nicholls *et al.*, 2014, Chalmers *et al.*, 2016). All these studies found similar reliability scores using 3D

models when compared with traditional plaster models. Use of 3D models also means that assessors do not need to be in the same physical location to score models, which is particularly convenient for intercentre studies (Fowler *et al.*, 2019). However, plaster models are still the dominant format and there are issues that need to be ironed out with 3D models before their wider adoption. One particular problem that has been highlighted is the error in correctly registering the intermaxillary relationship of models (Nicholls *et al.*, 2014), with one study reporting that 10% of models were incorrectly registered and needed digitally rearticulating prior to scoring (Fowler *et al.*, 2019). Correctly registered models are clearly crucial in occlusal scoring using any cleft indices.

It is possible that low and middle income countries may not have easy access to 3D models. For comparison of results it is important that an index is adopted and accessible worldwide. The results of this study demonstrate that use of reference models improves reliability when specialty trainees score models using the Modified 5-Year-Olds' Index. However, consultants and post-CCST trainees were able to use the Index reliably with only an information sheet and no reference models. The index can therefore be used by experienced operators to score study models in whichever format is locally available.

6.3 Reliability

Intraclass correlation coefficients were chosen for this study based on results generated in developing the Modified 5-Year-Olds' Index (Mittal *et al.*, 2018). In this latter study, examiners scored study models using both the categorical Modified 5-Year-Olds' Index and a VAS. The VAS scores were used to indicate the most appropriate weighting between categories. Linear regression models were fitted with the VAS and the Modified 5-Year-Olds'

Index score as either linear, quadratic or cubic exposure variable. Model comparison using likelihood ratios demonstrated strong evidence that quadratic modelling was the best fit. Quadratic weighted kappa values are equivalent to ICC (Fleiss and Cohen, 1973, Schuster, 2004) and as ICC was used for expert reliability assessment of the Modified 5-Year-Olds' Index (Mittal *et al.*, 2018) it was used in the study reported here for consistency.

The ICCs were calculated based on a two-way random effects model. Using this model means the results can be generalised to other raters possessing the same characteristics as the raters in this study *i.e.* trainee or qualified specialist orthodontists. High levels of inter- and intra-rater reliability have been demonstrated when the Modified 5-Year-Olds' Index was tested on expert examiners with extensive experience in the application of the 5-Year-Olds' Index (Mittal *et al.*, 2018).

6.3.1 Interpretation of ICCs

There are no rigidly accepted standard values used determine what is and what is not an acceptable degree of reliability using ICC, and as such, guidelines for interpretation of ICC agreement measures vary. The most commonly used interpretation is that of Cicchetti (1994), who suggests the following:

- 0.75 – 1.00 excellent
- 0.60 – 0.74 good
- 0.040 - 0.59 fair
- <0.40 poor

Koo and Li (2016) suggest a slightly more stringent interpretation:

- >0.90 excellent
- 0.75 – 0.90 good
- 0.50 – 0.75 moderate
- <0.50 poor

It is important not just to look at the ICC when determining the reliability, but also the 95% confidence interval, as this indicates a 95% chance the true ICC lies between the two values. For example, the intra-rater reliability score for the consultant in Group 1 was 0.93 (0.88, 0.96) and based on Koo and Li's interpretation of ICC it would be appropriate to conclude that the level of reliability is good to excellent, or just excellent using Cicchetti's guideline. However, it must be taken into consideration that ICCs are a relative measure based on variation within the sample being analysed. High levels of agreement scoring models with only a small degree of variation will produce a result with a low ICC; a good level of reliability may be shown despite low levels of agreement scoring a sample of models with a great degree variation between them (Bland and Altman, 1990). For example, ICC results when scoring study models can vary depending on the index used. VAS scores have a theoretical infinite number of possible scores as they are effectively classless. If a 100mm VAS was used and measured in 1mm increments, it would produce 100 classes. This gives a much greater degree of variation than using the Modified 5-Year-Olds' Index with 7 categories, and therefore would result in a higher ICC at the same level of agreement.

Measurement of reliability with ICC showed generally high levels of reliability when the specialty trainees of Group 1 (Assessors 3-5) are excluded, with intra-rater ICCs ranging from

0.80 (95% CI 0.67, 0.88) to 0.93 (95% CI 0.88, 0.96) and inter-rater ICCs from 0.80 (95% CI 0.65, 0.89) to 0.94 (95% CI 0.90, 0.96). At the highest end of the range the ICCs and 95% CIs are comparable to the expert assessors when first testing the index (Mittal *et al.*, 2018). At the lower end of this range the ICCs also demonstrate high levels of agreement, though the confidence intervals are wider.

Except for Assessor 1, the intra-rater ICCs for all assessors in this study are lower than those of the two experts when first testing the index (Mittal *et al.*, 2018). Both experts demonstrated very high levels of agreement with ICC and 95% CI of 0.92 (0.89, 0.94) and 0.93 (0.91, 0.95) when testing the Modified 5-Year-Olds' Index on the entire CCUK cohort of 198 models. Inter-rater reliability scores were also high, with ICC and 95% CI of 0.89 (0.86, 0.92) and 0.91 (0.89, 0.93). This very high level of reliability is likely explained by the fact the two examiners are very experienced in cleft care and use of the original 5-Year-Olds' Index, and played a significant part in developing the Modified 5-Year-Olds' Index.

Categories 3 to 6 of the modified index appear to be the most difficult categories to score reliably, with the most variation in scoring seen across these categories. The boundaries between categories are relatively subjective and there are a higher number of categories to distinguish between in the Modified 5-Year-Olds' Index compared with the original Index. All these categories are narrower than the categories of the original 5-Year-Olds' Index they were developed from, and it is therefore unsurprising they are more difficult to score reliably. In the development of the Modified 5-Year-Olds' Index, the examiners felt there were more occasions that warranted discussion to decide upon a consensus score

compared with the original index, as there were more categories and therefore more potential sources of disagreement (Mittal *et al.*, 2018).

6.3.2 Comparison of reliability

6.3.2.1 Cleft indices

It is important that reliability of the Modified 5-Year-Olds' Index (Mittal *et al.*, 2018) compares favourably with that of other dentoalveolar cleft indices for it to be considered for wider use. There have been several studies testing reliability of dentoalveolar cleft indices either as the primary aim or as part of a wider study.

The dentoalveolar relationships of 54 children born with UCLP in Western Australia (Johnson *et al.*, 2000) were assessed using the 5-Year-Olds' Index (Atack *et al.*, 1997a). Scoring was carried out by a very small number of assessors with one orthodontic specialty trainee and one senior registrar. The results were analysed using linear weighted kappa statistics, with intra-rater reliability stated as very good and inter-rater reliability good. However, it was not clear how they came to the conclusion that the reliability is 'good' and 'very good'. The accepted levels of agreement, as described by Landis and Koch (1977) are:

- 0.80-1.00 Almost perfect agreement
- 0.61-0.80 Substantial agreement
- 0.41-0.60 Moderate agreement
- 0.21-0.40 Fair agreement
- 0.01-0.20 Slight agreement
- <0 Poor agreement

The levels of both intra- and interrater agreement are lower than those of this study when excluding Assessors 3-5. The methods of calculating reliability are different between the two studies, with Johnson *et al*. (2000) using linear weighted kappa statistics (Cicchetti and Allison, 1971) and this study employing the intraclass correlation coefficient. Under the assumption that the spread of models across categories is broadly similar, both approaches produce similar levels of agreement (Mitani *et al.*, 2017). It is a reasonable assumption to make given that the studies are assessing similar populations – patients born with UCLP.

Most studies reporting the reliability of dentoalveolar indices use linear weighted kappa statistics. This makes the assumption that there is a linear relationship between each of the index categories, with a lack of data to suggest alternative weighting, such as quadratic or cubic (Jones *et al.*, 2014). Whilst this is a reasonable assumption, it may not be correct. VAS scoring data was used to determine the relationship between the categories of the Modified 5-Year-Olds' Index which demonstrated the relationship to be quadratic (Mittal *et al.*, 2018). Quadratic weighted kappa is directly equivalent to ICCs (Fleiss and Cohen, 1973). A score of 1 indicates perfect agreement, with 0 indicating no agreement.

The 5-Year-Olds' Index has proven to have high levels of reliability in larger studies. A large multi-centre study in the USA (Flinn *et al.*, 2006) reported very high inter-rater linear weighted kappa scores (0.937 – 0.965) and high inter-rater kappa scores (0.797 – 0.891). The intra-rater results must be interpreted with caution however as the models were scored on two consecutive days, not accounting for the possibility of recall bias. The levels of agreement are similar to those in this study (excluding Assessors 3-5).

A study comparing indices to measure the dentoalveolar outcome of primary cleft surgery (Jones *et al.*, 2016), using the 5-Year-Olds' Index, GOSLON, Modified Huddart and Bodenham (MHB), EUROCRAN and simple overjet measurement, found the MHB to be the most reliable of the five tested. Reliability was calculated using weighted kappa statistics, although there was no mention of the type of weighting used. The two examiners were consultant orthodontists with experience in cleft, but no experience scoring models, and neither were calibrated in any of the indices prior to scoring. The author states the MHB index had the best inter-examiner reliability with almost perfect agreement, appearing to use Cohen's (1960) interpretation of levels of agreement. However, Cohen's interpretation is often seen as too lenient for health-related studies as it suggests a score as low as 0.41 may be acceptable (McHugh, 2012). Regardless of the interpretation, the levels of agreement for all five indices are lower than the reliability of the Modified 5-Year-Olds' Index in this study, if excluding Assessors 3-5. Another study compared the 5-Year-Olds' and the MHB indices (Mikoya *et al.*, 2015); both were rated good or better using Altman's (1991) method of interpretation. This is again a particularly lenient interpretation of the values, as Altman suggests that Kappa values 0.2 – 0.4 are 'fair' – values in this range are unlikely to be considered reliable in medical research.

The GOSLON Yardstick (Mars *et al.*, 1987) is the most widely used outcome measure for assessment of the results of primary surgery in patients born with UCLP at the age of 10. Reliability testing of the Modified 5-Year-Olds' Index in this study demonstrates comparable or higher levels of agreement when compared with studies testing reliability of the GOSLON Yardstick (Hathaway *et al.*, 2011, Dogan *et al.*, 2012, Nicholls *et al.*, 2014, Zhu *et al.*, 2016). Of these, the Americleft (Hathaway *et al.*, 2011), which is the largest of the studies, was

conducted across five cleft centres in the USA. The authors state that intra- and inter-rater reliabilty were very good, based on the interpretation of agreement by Landis and Koch (1977). This is another overly lenient interpretation of kappa statistics to use in medical research, with similar cutoff levels to those of Altman (1991). The authors provide minimal reliability data, publishing only average kappa values across all raters with no confidence intervals, and do not mention whether the weighted kappa is linear or quadratic. A confidence interval indicates a range of possible values for the 'true' value of kappa within a given probability and should be provided along with the weighted kappa values (Sim and Wright, 2005). They conclude that the GOSLON Yardstick proved capable of discriminating between dental arch relationships across all the study centres, and although not primarily a reliability study, the reliability scores form an important part of the research and so more detailed data could have been be expected.

The GOSLON Yardstick and a 10cm VAS were compared using plaster and 3D models in a recent study in New Zealand (Fowler *et al.*, 2019). Both the index and the VAS were found to be reliable methods of measuring dentoalveolar relationships using both plaster and digital formats. The levels of reliability are, however, lower than those of the Modified 5-Year-Olds' Index in this study. Mittal (2018) found a 100mm VAS to have lower ICCs than the Modified 5-Year-Olds' Index, which is the opposite of the result expected as a result of the significantly higher number of categories in the VAS, concluding that the VAS was not as reliable as the Modified 5-Year-Olds' Index.

6.3.2.2 Non-cleft indices

The reliability of the Modified 5-Year-Olds' Index (Mittal *et al.*, 2018) compares well with other indices routinely used in the field of orthodontics. The Index of Orthodontic Treatment Need (IOTN) (Brook and Shaw, 1989) was developed as a tool to measure orthodontic treatment priority. It has been reported as the most frequently used index in high-impact scientific journals (Bellot-Arcis *et al.*, 2012, Taghavi Bayat *et al.*, 2017), and is used in the UK National Health Service (NHS) to determine eligibility for treatment based on the 'worst' aspect of the malocclusion. In development of the index the dental health component (DHC) was scored with a much higher level of reliability than the aesthetic component (AC), reflecting its more objective nature. Intra-examiner agreement ranged from a kappa value of 0.754 to 0.837 and inter-examiner kappa values from 0.731 to 0.797 (Brook and Shaw, 1989). A later study comparing occlusal indices demonstrated similar levels of intra- and inter-rater reliability for the IOTN (Ovsenik and Primozic, 2007). The reliability level of the Modified 5-Year-Olds' Index in this study is higher than that of the IOTN, a valid and reliable index used widely throughout the UK.

The Index of Orthognathic Functional Treatment Need (IOFTN) (Ireland *et al.*, 2014) was developed to aid prioritisation of severe malocclusions not suitable for orthodontic treatment alone and has been shown to have moderate to good interrater reliability and good interrater reliability based on Cohen's (1960) interpretation of kappa statistics. As with the Modified 5-Year-Olds' Index, the IOFTN is said to have good face validity. A service evaluation of the IOFTN (Howard-Bowles *et al.*, 2017) found it simple to use and concluded that it was reliable. The reliability scores in both development and service evaluation of the IOFTN are lower than the reliability level of the Modified 5-Year-Olds' Index in this study.

The Peer Assessment Rating (PAR) Index (Richmond *et al.*, 1992) was developed to provide a single score calculated by summing all the occlusal anomalies found in the malocclusion of an individual, demonstrating how significantly it deviates from normal occlusion. This is calculated both before and after treatment to measure the degree of improvement. It has proven to be valid and reliable (Richmond *et al.*, 1992) and is used to monitor treatment outcome in the NHS as a contractual requirement. Excellent intra-examiner and inter-examiner reliability was reported in development of the index, with ICCs of >0.95 and 0.91 respectively (Richmond *et al.*, 1992). A study involving 10 examiners PAR scoring 206 models reported kappa scores for intra-examiner reliability of 0.877 and inter-examiner reliability of 0.831 (Pangrazio-Kulbersh *et al.*, 1999). The reliability scores for the Modified 5-Year-Olds' Index in this study compare favourably with the latter study and are not significantly lower than those of the former. The PAR Index has detailed instructions on scoring and is largely objective to use; the comparable levels of reliability with the PAR Index suggests that the verbal descriptors of the Modified 5-Year-Olds' Index enable clear differentiation between the categories for reliable scoring.

6.4 Calibration

Calibration is the process of checking, by comparison with a standard, the accuracy of a measuring instrument of any type. In the case of indices, this means ensuring that examiners are trained to score reliably when compared with a gold standard, such as the consensus scores of experts. Calibration is a pre-requisite for using the original 5-Year-Olds' Index (Atack *et al.*, 1997b). The GOSLON Yardstick also requires calibration prior to competent use, and regular recalibration is recommended to ensure continued accurate use

(Mossey *et al.*, 2003). Within the field of orthodontics, calibration is often necessary for some of the more complex indices. A study comparing examiners use of the Index of Complexity, Outcome and Need (ICON) for determining treatment need found higher levels of reliability for a calibrated orthodontist when compared with non-calibrated orthodontists (Louwerse *et al.*, 2006). The Peer Assessment Rating (PAR) Index was designed for use by calibrated examiners (Richmond *et al.*, 1992). Some of the original authors of the PAR Index also demonstrated that rigorous calibration was sufficient to teach non-dental staff to use the index with a high degree of reliability (Richmond *et al.*, 1993). Calibrating non-dental staff to use an index reliably frees up valuable clinical time and should improve its objective and impartial use.

A study comparing different indices for dentoalveolar outcomes of primary cleft surgery, without prior calibration in any of the indices, found that the assessors felt they would have improved their reliability using the 5-Year-Olds' Index if they had been calibrated first (Jones *et al.*, 2016). A recent study into the dentoalveolar outcome of primary surgery in Sweden using the quality registry for CLP scored models using the MHB and 5-Year Olds' indices (Pegelow *et al.*, 2020). They noted an obvious difference in scores using the 5-Year-Olds' Index between one examiner that had not been previously calibrated versus thirteen that had prior calibration and experience using the index. The scoring difference must be interpreted with caution however, as comparing one inexperienced examiner with thirteen experienced examiners is clearly too unbalanced to draw any strong conclusion. All cleft centres in Sweden plan to calibrate examiners annually to maintain reliability and validity of scoring.

As the 5-Year-Olds' Index requires calibration, it was expected that the assessors calibrated in use of the Modified 5-Year-Olds' Index would produce the highest reliability scores. However, this was not the case in this study. All assessors except the specialty trainees of Group 1 demonstrated high levels of reliability. This suggests that the verbal descriptors generated for each of the categories are comprehensive and sufficiently clear to enable accurate use of the index without calibration. As an example, scoring a model with an edge-to-edge incisor relationship and no buccal crossbites using the 5-Year-Olds' Index could be confusing, in that the model would be scored in category 3 with respect to the overjet, but category 1 when considering the transverse relationship. Once calibrated, assessors have learnt the convention that this model would be scored in category 2 (Atack, 2012), but this is not stated in the published index. In the Modified 5-Year-Olds' Index the above model would be scored in category 3, with ambiguity removed from the index. The ability to use an index reliably without calibration is an advantage, and could be a particularly important feature for use in the developing world (Dobbyn *et al.*, 2012).

6.5 Validity

Validity is the extent to which the scores from a measure represent the variable they are intended to (Steiner *et al.*, 2008). True validation of the Modified 5-Year-Olds' Index is not possible, as it would require comparing individuals born with UCLP on the cleft surgical pathway with those born with UCLP who had no treatment at all after primary surgery. This is not possible as all children born with UCLP in the developed world receive all treatment necessary and withholding treatment based on research is clearly unethical. The outcome of primary surgery becomes distorted over time with any further treatment(s) such as secondary alveolar bone grafting, and the growth pattern of the individual.

Attempts to assess the validity of the original 5-Year-Olds' Index have been made. It was compared with the GOSLON Yardstick for predicting long-term relationships (Mars *et al.*, 2006) with the latter being found more reliable, though it was recommended changes were made to both indices. It was noted that the 5-Year-Olds' Index was developed using a sample of models with overjets no greater than 2mm, when the Gothenburg series of models by contrast commonly demonstrated overjets of greater than 5mm. It was felt that there was a significant difference between these overjets, but they would both be categorised in Category 1 of the 5-Year-Olds' Index. Whilst overjet and AP relationship are of major importance, the scope to recognise such differences in overjets has not changed in the Modified 5-Year-Olds' Index. As a method of establishing validity in comparing cleft indices, Jones *et al*. (2016) designated the 5-Year-Olds' Index the gold standard with which to compare the other indices at the age of 5 years, and the GOSLON Yardstick the gold standard for comparison at 10 years of age. Whilst this is a reasonable method, there is no information available to assess the validity of the designated gold standard indices and so it is not possible to establish true validity.

Instead, the 5-Year-Olds' Index is stated as having face validity (Atack *et al.*, 1997b), in that it subjectively appears to measure what it sets out to measure as deemed by experts in the field (Mosier, 1947). The Modified 5-Year-Olds' Index was found to be valid in comparison to the 5-Year-Olds' Index when applied to models from the CCUK cohort, with a more even distribution of scores across all of the categories and higher levels of reliability in model scoring (Mittal *et al.*, 2018).

6.6 Strengths of the study

The study was conducted in a methodical manner. The sample size of 50 models was determined by a statistical calculation based on confidence intervals. The models were selected from the CCUK total model set to best reflect the spread of dentoalveolar outcomes from the cohort.

The calibration course was delivered by NEA, an expert who developed the 5-Year-Olds' Index and was also instrumental in development of the Modified 5-Year-Olds' Index. This ensured the assessors who were calibrated benefited from the highest level of knowledge and experience, with the opportunity to discuss and clarify the use of the index prior to model scoring.

The study was designed to reduce any bias. Assessors with minimal or no experience in treating patients born with a cleft were chosen to reduce any systematic bias, as experience in cleft care is a significant factor in assessing the outcome of treatment. A washout period of at least four weeks between scoring sessions minimised the effect of memory bias, although the likelihood of participants being able to remember the score they gave at an even shorter interval is low (Fowler *et al.*, 2019).

The Index was tested on enough participants across the groups to enable confidence in the reliability scores. Equal numbers of assessors at each level of training were selected to ensure even experience levels across the groups to enable comparison.

6.7 Weaknesses of the study

One of the consultant assessors (Assessor 1) participating in this study has previously carried out research looking at dentoalveolar cleft indices (Jones *et al.*, 2016). The research involved arranging for two consultant examiners to score study models using five different cleft indices in order to determine which should be used to measure primary surgical outcome for UCLP patients. Whilst he did not score study models or participate in any calibration exercises in his own research, he could be considered to have more experience than the consultants in the other groups in this study. This could therefore be a potential source of bias and may have contributed towards his high reliability scores.

7.0 CONCLUSIONS

- The Modified 5-Year-Olds' Index was found to be a reliable method of measuring the dentoalveolar outcome of primary cleft surgery at 5 years of age when model scoring is carried out by consultants or post-CCST trainees with no previous experience of cleft scoring, and only an information sheet as a guide.
- Less experienced orthodontic specialty trainees were able to use the Modified 5-Year-Olds' Index reliably, but performed this best when provided with a set of reference models in addition to an information sheet.
- Calibration in use of the Modified 5-Year-Olds' Index does not improve reliability compared to being provided with an information sheet and reference models alone.
- For ease of use and uptake across cleft departments internationally it can be recommended that the index be used for scoring study models by consultants or higher level specialty trainees without calibration or access to reference models.

8.0 FUTURE RESEARCH

8.1 Investigation of the importance of orthodontic training on reliable use of the Modified 5-Year-Olds' Index

The Modified 5-Year-Olds' Index was shown to be reliable when scoring was undertaken by assessors who all had experience in orthodontics, at specialty trainee level or higher. Testing and comparing reliability of the Modified 5-Year-Olds' Index when used by non-orthodontically trained assessors would be beneficial. Increasing the number of individuals able to apply the indices would be more economical on clinical time and should improve the objectivity and impartiality of the index.

8.2 Investigation into the reason the middle categories of the Modified 5-Year-Olds' Index are more difficult to score reliably

The majority of both intra- and inter-examiner disagreements occurred between categories 3 and 6. It would be useful to investigate what makes these categories hard to agree on. The information gained from this research could be used to adapt the verbal descriptors of the categories and/or the information sheet to provide further clarity for more reliable assessment of study models.

8.3 Investigation into the reliability of digital versus plaster study models in use of the Modified 5-Year-Olds' Index

Whilst three-dimensional digital study models have been shown to be reliable in comparison to plaster models in a number of studies, the Modified 5-Year-Olds' Index has only been tested using plaster models. There are a number of previously discussed

advantages of digital study models, and they are likely to become increasingly used within orthodontics in the future. It would be useful to determine whether assessment of digital study models using the Modified 5-Year-Olds' Index proves to be reliable in comparison to plaster models.

9.0 REFERENCES

AKRAM, A., MCKNIGHT, M. M., BELLARDIE, H., BEALE, V. & EVANS, R. D. 2015. Craniofacial malformations and the orthodontist. Br Dent J*,* 218**,** 129-141.

AL-GHATAM, R., JONES, T. E. M., IRELAND, A. J., ATACK, N. E., CHAWLA, O., DEACON, S., ALBERY, L., COBB, A. R. M., CADOGAN, J., LEARY, S., WAYLEN, A., WILLS, A. K., RICHARD, B., BELLA, H., NESS, A. R. & SANDY, J. R. 2015. Structural outcomes in the Cleft Care UK study. Part 2: Dento-facial outcomes. Orthodontics and Craniofacial Research, 18, 14-24.

ALTALIBI, M., SALTAJI, H., EDWARDS, R., MAJOR, P. W. & FLORES-MIR, C. 2013. Indices to assess malocclusions in patients with cleft lip and palate. Eur J Orthod, 35, 772-82.

ALTMAN, D. G. 1991. Practical statistics for medical research, Chapman and Hall.

AMANAT, N. & LANGDON, J. D. 1991. Secondary alveolar bone grafting in clefts of the lip and palate. J Craniomaxillofac Surg, 19, 7-14.

ANDERSON, O., NI, Z., MOLLER, H., COUPLAND, V. H., DAVIES, E. A., ALLUM, W. H. & HANNA, G. B. 2011. Hospital volume and survival in oesophagectomy and gastrectomy for cancer. Eur J Cancer, 47, 2408-14.

ASHER-MCDADE, C., BRATTSTROM, V., DAHL, E., MCWILLIAM, J., MOLSTED, K., PLINT, D. A., PRAHL-ANDERSEN, B., SEMB, G. & SHAW, W. C. 1992. A six-center international study of treatment outcome in patients with clefts of the lip and palate: Part 4. Assessment of nasolabial appearance. Cleft Palate Craniofac J, 29, 409-12.

ASHER-MCDADE, C., ROBERTS, C., SHAW, W. C. & GALLAGER, C. 1991. Development of a method for rating nasolabial appearance in patients with clefts of the lip and palate. Cleft Palate Craniofac J, 28, 385-90

ASSUNCAO, A. G. 1992. The V.L.S. classification for secondary deformities in the unilateral cleft lip: clinical application. Br J Plast Surg, 45, 293-6.

ATACK, N. 2012. 5 Year Olds' Index Calibration Handbook (Japan).

ATACK, N., HATHORN, I., MARS, M. & SANDY, J. 1997a. Study models of 5 year old children as predictors of surgical outcome in unilateral cleft lip and palate. Eur J Orthod, 19, 165-170.

ATACK, N. E., HATHORN, I. S., SEMB, G., DOWELL, T. & SANDY, J. 1997b. A New Index for Assessing Surgical Outcome in Unilateral Cleft Lip and Palate Subjects Aged Five: Reproducibility and Validity. Cleft Palate-Craniofac J, 34(3), 242-6

BEARN, D., MILDINHALL, S., MURPHY, T., MURRAY, J. J., SELL, D., SHAW, W. C., WILLIAMS, A. C. & SANDY, J. R. 2001. Cleft lip and palate care in the United Kingdom--the Clinical Standards Advisory Group (CSAG) Study. Part 4: outcome comparisons, training, and conclusions. Cleft Palate Craniofac J, 38, 38-43.

BELLOT-ARCIS, C., MONTIEL-COMPANY, J. M., ALMERICH-SILLA, J. M., PAREDES-GALLARDO, V. & GANDIA-FRANCO, J. L. 2012. The use of occlusal indices in high-impact literature. Community Dent Health, 29, 45-8.

BERGLAND, O., SEMB, G. & ABYHOLM, F. E. 1986. Elimination of the residual alveolar cleft by secondary bone grafting and subsequent orthodontic treatment. Cleft Palate J, 23, 175-205.

BLAND, J. M. & ALTMAN, D. G. 1990. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. Comput Biol Med, 20, 337-40.

BROOK, P. H. & SHAW, W. C. 1989. The development of an index of orthodontic treatment priority. Eur J Orthod, 11, 309-20.

BUJ-ACOSTA, C., PAREDES-GALLARDO, V., MONTIEL-COMPANY, J. M., ALBALADEJO, A. & BELLOT-ARCIS, C. 2017. Predictive validity of the GOSLON Yardstick index in patients with unilateral cleft lip and palate: A systematic review. PLoS One, 12, e0178497.

CALZOLARI, E., PIERINI, A., ASTOLFI, G., BIANCHI, F., NEVILLE, A. J. & RIVIERI, F. 2007. Associated anomalies in multi-malformed infants with cleft lip and palate: An epidemiologic study of nearly 6 million births in 23 EUROCAT registries. Am J Med Genet Part A, 143A, 528-537.

CHALMERS, E. V., MCINTYRE, G. T., WANG, W., GILLGRASS, T., MARTIN, C. B. & MOSSEY, P. A. 2016. Intraoral 3D Scanning or Dental Impressions for the Assessment of Dental Arch Relationships in Cleft Care: Which is Superior? Cleft Palate Craniofac J, 53, 568-77.

CHAPMAN, K. L., HARDIN-JONES, M. A., GOLDSTEIN, J. A., HALTER, K. A., HAVLIK, R. J. & SCHULTE, J. 2008. Timing of palatal surgery and speech outcome. Cleft Palate Craniofac J, 45, 297-308.

CHAWLA, O., ATACK, N. E., DEACON, S. A., LEARY, S. D., IRELAND, A. J. & SANDY, J. R. 2013. Three-dimensional digital models for rating dental arch relationships in unilateral cleft lip and palate. Cleft Palate Craniofac J, 50(2),182-6

CICCHETTI, D. 1994. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instrument in Psychology. Psychological Assessment, 6, 284-290.

CICCHETTI, D. V. & ALLISON, T. 1971. A New Procedure for Assessing Reliability of Scoring EEG Sleep Recordings. American Journal of EEG Technology, 11, 101-110.

COHEN, J. 1960. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20, 37-46.

DE MULDER, D., CADENAS DE LLANO-PERULA, M., JACOBS, R., VERDONCK, A. & WILLEMS, G. 2018. Three-dimensional radiological evaluation of secondary alveolar bone grafting in cleft lip and palate patients: a systematic review. Dentomaxillofac Radiol, 20180047.

DIKKEN, J. L., DASSEN, A. E., LEMMENS, V. E., PUTTER, H., KRIJNEN, P., VAN DER GEEST, L., BOSSCHA, K., VERHEIJ, M., VAN DE VELDE, C. J. & WOUTERS, M. W. 2012. Effect of hospital volume on postoperative mortality and survival after oesophageal and gastric cancer surgery in the Netherlands between 1989 and 2009. Eur J Cancer, 48, 1004-13.

DIXON, M. J., MARAZITA, M. L., BEATY, T. H. & MURRAY, J. C. 2011. Cleft lip and palate: understanding genetic and environmental influences. Nature Reviews Genetics, 12, 167-178.

DOBBYN, L. M., WEIR, J. T., MACFARLANE, T. V. & MOSSEY, P. A. 2012. Calibration of the modified Huddart and Bodenham scoring system against the GOSLON/5-year-olds' index for unilateral cleft lip and palate. Eur J Orthod, 34, 762-767.

DOGAN, S., OLMEZ, S. & SEMB, G. 2012. Comparative assessment of dental arch relationships using Goslon Yardstick in patients with unilateral complete cleft lip and palate using dental casts, two-dimensional photos, and three-dimensional images. Cleft Palate Craniofac J, 49, 347-51.

DOROS, G. & LEW, R. 2010. Design based on intra-class correlation coefficients. American Journal of Biostatistics, 1,1.

ECKSTEIN, D. A., WU, R. L., AKINBIYI, T., SILVER, L. & TAUB, P. J. 2011. Measuring quality of life in cleft lip and palate patients: currently available patient-reported outcomes measures. Plast Reconstr Surg, 128, 518e-526e.

EDLER, R., RAHIM, M. A., WERTHEIM, D. & GREENHILL, D. 2010. The use of facial anthropometrics in aesthetic assessment. Cleft Palate Craniofac J, 47, 48-57.

ENEMARK, H., SINDET-PEDERSEN, S., BUNDGAARD, M. & SIMONSEN, E. K. 1988. Combined orthodontic-surgical treatment of alveolar clefts. Ann Plast Surg, 21, 127-33.

FITZSIMONS, K. J., MUKARRAM, S., COPLEY, L. P., DEACON, S. A. & VAN DER MEULEN, J. H. 2012. Centralisation of services for children with cleft lip or palate in England: a study of hospital episode statistics. BMC Health Serv Res, 12, 148.

FLEISS, J. L. & COHEN, J. 1973. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. Educational and Psychological Measurement, 33, 613-619.

FLEMING, P. S., MARINHO, V. & JOHAL, A. 2011. Orthodontic measurements on digital study models compared with plaster models: A systematic review. Orthodontics and Craniofacial Research, 14, 1-16.

FLINN, W., LONG, R. E., GARATTINI, G. & SEMB, G. 2006. A Multicenter Outcomes Assessment of Five-Year-Old Patients With Unilateral Cleft Lip and Palate. Cleft Palate Craniofac J, 43, 253-253.

FOWLER, P., BELLARDIE, H., SHAW, B., EYRES, P., SEMB, G. & THOMPSON, J. 2019. Reliability of a Categorical Scale (GOSLON) and a Continuous Scale (10-cm Visual Analog Scale) for Assessing Dental Arch Relationships Using Conventional Plaster and 3D Digital Orthodontic Study Models of Children With Complete Unilateral Cleft Lip and Palate. Cleft Palate Craniofac J, 56, 84-89.

FOWLER, P. V., AL-ANI, A. H. & THOMPSON, J. M. D. 2018. Comparison of Reliability of Categorical and Continuous Scales for Radiographic Assessments of Bone Infill Following Secondary Alveolar Bone Grafting. Cleft Palate Craniofac J, 55, 269-275.

FUDALEJ, P., KATSAROS, C., BONGAARTS, C., DUDKIEWICZ, Z. & KUIJPERS-JAGTMAN, A. M. 2011. Dental arch relationship in children with complete unilateral cleft lip and palate following one-stage and three-stage surgical protocols. Clin Oral Investig, 15, 503-10.

GORLIN, R. J., COHEN JR, M. M. & HENNEKAM, R. C. M. 2001. Syndromes of the head and neck, Oxford University Press.

GRAY, D. & MOSSEY, P. A. 2005. Evaluation of a modified Huddart/Bodenham scoring system for assessment of maxillary arch constriction in unilateral cleft lip and palate subjects. Eur J Orthod, 27, 507-11.

GUO, J., LI, C., ZHANG, Q., WU, G., DEACON, S. A., CHEN, J., HU, H., ZOU, S. & YE, Q. 2011. Secondary bone grafting for alveolar cleft in children with cleft lip or cleft lip and palate. Cochrane Database of Systematic Reviews.

HATHAWAY, R., DASKALOGIANNAKIS, J., MERCADO, A., RUSSELL, K., LONG, R. E., JR., COHEN, M., SEMB, G. & SHAW, W. 2011. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 2. Dental arch relationships. Cleft Palate Craniofac J, 48, 244-51.

HATHORN, I. S., ATACK, N. E., BUTCHER, G., DICKSON, J., DURNING, P., HAMMOND, M., KNIGHT, H., MITCHELL, N., NIXON, F., SHINN, D. & SANDY, J. R. 2006. Centralization of services: standard setting and outcomes. Cleft Palate Craniofac J, 43, 401-5.

HEIDBUCHEL, K. L. & KUIJPERS-JAGTMAN, A. M. 1997. Maxillary and mandibular dental-arch dimensions and occlusion in bilateral cleft lip and palate patients from 3 to 17 years of age. Cleft Palate Craniofac J, 34, 21-6.

HODGKINSON, P. D., BROWN, S., DUNCAN, D., GRANT, C., MCNAUGHTON, A. M. Y., THOMAS, P. & MATTICK, C. R. 2005. Management of children with cleft lip and palate: A review describing the application of multidisciplinary team working in this condition based upon the experiences of a regional cleft lip and palate centre in the United Kingdom. Fetal and Maternal Medicine Review, 16, 1-27.

HOWARD-BOWLES, E., HO, A. Y. J., ULHAQ, A. & MCGUINNESS, N. J. P. 2017. The application of the Index of Orthognathic Functional Treatment Need (IOFTN): service evaluation and impact. J Orthod, 44, 97-104.

HUDDART, A. G. & BODENHAM, R. S. 1972. The evaluation of arch form and occlusion in unilateral cleft palate subjects. Cleft Palate J, 9, 194-209.

HUNT, O., BURDEN, D., HEPPER, P. & JOHNSTON, C. 2005. The psychosocial effects of cleft lip and palate: a systematic review. Eur J Orthod, 27, 274-85.

IRELAND, A. J., CUNNINGHAM, S. J., PETRIE, A., COBOURNE, M. T., ACHARYA, P., SANDY, J. R. & HUNT, N. P. 2014. An Index of Orthognathic Functional Treatment Need (IOFTN). J Orthod. 41(2):77-83.

JOHN, A., SELL, D., SWEENEY, T., HARDING-BELL, A. & WILLIAMS, A. 2006. The cleft audit protocol for speech-augmented: A validated and reliable measure for auditing cleft speech. Cleft Palate Craniofac J, 43, 272-88.

JOHNSON, C. Y. & LITTLE, J. 2008. Folate intake, markers of folate status and oral clefts: is the evidence converging? Int J Epidemiol, 37, 1041-58.

JOHNSON, N. & SANDY, J. 2003. An aesthetic index for evaluation of cleft repair. Eur J Orthod, 25, 243-9.

JOHNSON, N., WILLIAMS, A. C., SINGER, S., SOUTHALL, P., ATACK, N. & SANDY, J. R. 2000. Dentoalveolar relations in children born with a unilateral cleft lip and palate (UCLP) in Western Australia. The Cleft Palate-Craniofacial Journal, 37, 12-16.

JONES, C. M., ROTH, B., MERCADO, A. M., RUSSELL, K. A., DASKALOGIANNAKIS, J., SAMSON, T. D., HATHAWAY, R. R., SMITH, A., MACKAY, D. R. & LONG, R. E., JR. 2018. The Americleft Project: Comparison of Ratings Using Two-Dimensional Versus Three-Dimensional Images for Evaluation of Nasolabial Appearance in Patients With Unilateral Cleft Lip and Palate. J Craniofac Surg, 29, 105-108.

JONES, T., AL-GHATAM, R., ATACK, N., DEACON, S., POWER, R., ALBERY, L., IRELAND, T. & SANDY, J. 2014. A review of outcome measures used in cleft care. J Orthod, 41, 128-40.

JONES, T., LEARY, S., ATACK, N., IRELAND, A. & SANDY, J. 2016. Which index should be used to measure primary surgical outcome for unilateral cleft lip and palate patients? Eur J Orthod, 38(4), 345-352.

KAPPEN, I., BITTERMANN, D., JANSSEN, L., BITTERMANN, G. K. P., BOONACKER, C., HAVERKAMP, S., DE WILDE, H., VAN DER HEUL, M., SPECKEN, T. F., KOOLE, R., KON, M., BREUGEM, C. C. & MINK VAN DER MOLEN, A. B. 2017. Long-Term Follow-Up Study of Young Adults Treated for Unilateral Complete Cleft Lip, Alveolus, and Palate by a Treatment Protocol Including Two-Stage Palatoplasty: Speech Outcomes. Arch Plast Surg, 44, 202-9.

KAUFMAN, Y., BUCHANAN, E. P., WOLFSWINKEL, E. M., WEATHERS, W. M. & STAL, S. 2012. Cleft Nasal Deformity and Rhinoplasty. Semin Plast Surg. 26(04), 184-190

KINDELAN, J. D., NASHED, R. R. & BROMIGE, M. R. 1997. Radiographic assessment of secondary autogenous alveolar bone grafting in cleft lip and palate patients. Cleft Palate Craniofac J, 34, 195-8.

KIRBSCHUS, A., GESCH, D., HEINRICH, A. & GEDRANGE, T. 2006. Presurgical nasoalveolar molding in patients with unilateral clefts of lip, alveolus and palate. Case study and review of the literature. Journal of Cranio-Maxillofacial Surgery, 34, 45-48.

KLASSEN, A. F., TSANGARIS, E., FORREST, C. R., WONG, K. W., PUSIC, A. L., CANO, S. J., SYED, I., DUA, M., KAINTH, S., JOHNSON, J. & GOODACRE, T. 2012. Quality of life of children treated for cleft lip and/or palate: a systematic review. J Plast Reconstr Aesthet Surg, 65, 547-57.

KOO, T. K. & LI, M. Y. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of chiropractic medicine, 15, 155-163.

KORNBLUTH, M., CAMPBELL, R. E., DASKALOGIANNAKIS, J., ROSS, E. J., GLICK, P. H., RUSSELL, K. A., DOUCET, J.-C., HATHAWAY, R. R., LONG, R. E. & SITZMAN, T. J. 2018. Active Presurgical Infant Orthopedics for Unilateral Cleft Lip and Palate: Intercenter Outcome Comparison of Latham, Modified McNeil, and Nasoalveolar Molding. Cleft Palate Craniofac J, 55, 639-648.

KRAMER, F. J., GRUBER, R., FIALKA, F., SINIKOVIC, B., HAHN, W. & SCHLIEPHAKE, H. 2009. Quality of life in school-age children with orofacial clefts and their families. J Craniofac Surg, 20, 2061-6.

KRIENS, O. 1989. LAHSHAL-A Concise Documentation System for Cleft Lip, Alveolus and Palate Diagnoses What is a Cleft Lip and Palate?: A Multidisciplinary Update, 32-3

KUIJPERS-JAGTMAN, A. M., NOLLET, P. J., SEMB, G., BRONKHORST, E. M., SHAW, W. C. & KATSAROS, C. 2009. Reference photographs for nasolabial appearance rating in unilateral cleft lip and palate. J Craniofac Surg, 20 Suppl 2, 1683-6.

LALKHEN, A. G. & MCCLUSKEY, A. 2008. CLINICAL TESTS: SENSITIVITY AND SPECIFICITY. Continuing Education in Anaesthesia Critical Care & Pain, 8, 221-223.

LANDIS, J. R. & KOCH, G. G. 1977. The measurement of observer agreement for categorical data. Biometrics, 33, 159-74.

LITTLE, J., GILMOUR, M., MOSSEY, P. A., FITZPATRICK, D., CARDY, A., CLAYTON-SMITH, J. & FRYER, A. E. 2008. Folate and clefts of the lip and palate--a U.K.-based case-control study: Part I: Dietary and supplemental folate. Cleft Palate Craniofac J, 45, 420-7.

LONG, R. E., JR., HATHAWAY, R., DASKALOGIANNAKIS, J., MERCADO, A., RUSSELL, K., COHEN, M., SEMB, G. & SHAW, W. 2011. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 1. Principles and study design. Cleft Palate Craniofac J, 48, 239-43.

LOUWERSE, T. J., AARTMAN, I. H., KRAMER, G. J. & PRAHL-ANDERSEN, B. 2006. The reliability and validity of the Index of Complexity, Outcome and Need for determining treatment need in Dutch orthodontic practice. Eur J Orthod, 28, 58-64.

MAARSE, W., BERGE, S. J., PISTORIUS, L., VAN BARNEVELD, T., KON, M., BREUGEM, C. & MINK VAN DER MOLEN, A. B. 2010. Diagnostic accuracy of transabdominal ultrasound in detecting prenatal cleft lip and palate: a systematic review. Ultrasound Obstet Gynecol, 35, 495-502.

MARS, M., ASHER-MCDADE, C., BRATTSTRÖM, V., DAHL, E., MCWILLIAM, J., MØLSTED, K., A. PLINT, D., PRAHL-ANDERSEN, B., SEMB, G. & C. SHAW, W. 1992. A Six-Center International Study of Treatment Outcome in Patients with Clefts of the Lip and Palate: Part 3. Dental Arch Relationships. Cleft Palate Craniofac J, 29, 405-8.

MARS, M., BATRA, P. & WORRELL, E. 2006. Complete unilateral cleft lip and palate: validity of the five-year index and the Goslon yardstick in predicting long-term dental arch relationships. Cleft Palate Craniofac J, 43, 557-62.

MARS, M., PLINT, D. A., HOUSTON, W. J., BERGLAND, O. & SEMB, G. 1987. The Goslon Yardstick: a new system of assessing dental arch relationships in children with unilateral clefts of the lip and palate. Cleft Palate J, 24, 314-22.

MCBRIDE, W. A., MOSSEY, P. A. & MCINTYRE, G. T. 2013. Reliability, completeness and accuracy of cleft subphenotyping as recorded on the CLEFTSiS (Cleft Service in Scotland) electronic patient record. Surgeon, 11, 313-8.

MCGRAW, K. & WONG, S. P. 1996. Forming Inferences About Some Intraclass Correlation Coefficients. Psychological Methods, 1, 30-46.

MCHUGH, M. L. 2012. Interrater reliability: the kappa statistic. Biochem Med (Zagreb).

MEDINA, J., FITZSIMONS, K., DEACON, S., RUSSELL, C., WAHEDALLY, H., PARK, M. H. & VAN DER MEULEN, J. 2019. CRANE 2019 Annual Report.

MERCADO, A., RUSSELL, K., HATHAWAY, R., DASKALOGIANNAKIS, J., SADEK, H., LONG, R. E., JR., COHEN, M., SEMB, G. & SHAW, W. 2011. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 4. Nasolabial aesthetics. Cleft Palate Craniofac J, 48, 259-64.

MERCADO, A. M., RUSSELL, K. A., DASKALOGIANNAKIS, J., HATHAWAY, R. R., SEMB, G., OZAWA, T., SMITH, A., LIN, A. Y. & LONG, R. E., JR. 2016. The Americleft Project: A Proposed

Expanded Nasolabial Appearance Yardstick for 5- to 7-Year-Old Patients With Complete Unilateral Cleft Lip and Palate (CUCLP). Cleft Palate Craniofac J, 53, 30-7.

MEYER, K., WERLER, M., HAYES, C. & MITCHELL, A. 2003. Low maternal alcohol consumption during pregnancy and oral clefts in offspring: The Slone birth defects study. Birth Defects Research Part A: Clinical and Molecular Teratology, 67, 509-514.

MEYER-MARCOTTY, P., GERDES, A. B., REUTHER, T., STELLZIG-EISENHAUER, A. & ALPERS, G. W. 2010. Persons with cleft lip and palate are looked at differently. J Dent Res, 89, 400-4.

MEYER-MARCOTTY, P. & STELLZIG-EISENHAUER, A. 2009. Dentofacial self-perception and social perception of adults with unilateral cleft lip and palate. J Orofac Orthop, 70, 224-36.

MIKOYA, T., SHIBUKAWA, T., SUSAMI, T., SATO, Y., TENGAN, T., KATASHIMA, H., OYAMA, A., MATSUZAWA, Y., ITO, Y. & FUNAYAMA, E. 2015. Dental arch relationship outcomes in one- and two-stage palatoplasty for Japanese patients with complete unilateral cleft lip and palate. Cleft Palate Craniofac J, 52, 277-86.

MITANI, A. A., FREER, P. E. & NELSON, K. P. 2017. Summary measures of agreement and association between many raters' ordinal classifications. Ann Epidemiol, 27, 677-685.

MITCHELL, L. E. & RISCH, N. 1992. Mode of inheritance of nonsyndromic cleft lip with or without cleft palate: a reanalysis. Am J Hum Genet, 51, 323-32.

MITTAL, T. K., IRELAND, A. J., ATACK, N. E., LEARY, S. D., RUSSELL, J. I., DEACON, S. A., NESS, A. R. & SANDY, J. R. 2018. Outcome Measures in UCLP: The Modified 5-Year-Olds'-Index—Development and Reliability. Cleft Palate Craniofac J, 56, 248-256.

MORRIS, T., ROBERTS, C. & SHAW, W. C. 1994. Incisal overjet as an outcome measure in unilateral cleft lip and palate management. Cleft Palate Craniofac J, 31, 142-5.

MOSIER, C. I. 1947. A Critical Examination of the Concepts of Face Validity. Educational and Psychological Measurement, 7, 191-205.

MOSMULLER, D. G., BIJNEN, C. L., KRAMER, G. J., DISSE, M. A., PRAHL, C., KUIK, D. J., NIESSEN, F. B. & DON GRIOT, J. P. 2015. The Asher-McDade Aesthetic Index in Comparison With Two Scoring Systems in Nonsyndromic Complete Unilateral Cleft Lip and Palate Patients. J Craniofac Surg, 26, 1242-5.

MOSMULLER, D. G. M., MENNES, L. M., PRAHL, C., KRAMER, G. J. C., DISSE, M. A., VAN COUWELAAR, G. M., NIESSEN, F. B. & GRIOT, J. 2017. The Development of the Cleft Aesthetic Rating Scale: A New Rating Scale for the Assessment of Nasolabial Appearance in Complete Unilateral Cleft Lip and Palate Patients. Cleft Palate Craniofac J, 54, 555-561.

MOSSEY, P. A., CLARK, J. D. & GRAY, D. 2003. Preliminary investigation of a modified Huddart/Bodenham scoring system for assessment of maxillary arch constriction in unilateral cleft lip and palate subjects. Eur J Orthod, 25, 251-7.

MOSSEY, P. A., LITTLE, J., MUNGER, R. G., DIXON, M. J. & SHAW, W. C. 2009. Cleft lip and palate. The Lancet, 374, 1773-1785.

MUNGER, R. G., ROMITTI, P. A., DAACK-HIRSCH, S., BURNS, T. L., MURRAY, J. C. & HANSON, J. 1996. Maternal alcohol use and risk of orofacial cleft birth defects. Teratology, 54, 27-33.

MURRAY, J. J. 2003. Cleft lip and palate services. A review of developments five years after the CSAG report. Int J Paediatr Dent, 13, 395-403

MURRAY, L., ARTECHE, A., BINGLEY, C., HENTGES, F., BISHOP, D. V., DALTON, L., GOODACRE, T. & HILL, J. 2010. The effect of cleft lip on socio-emotional functioning in school-aged children. J Child Psychol Psychiatry, 51, 94-103.

MØLSTED, K., ASHER-MCDADE, C., BRATTSTRÖM, V., DAHL, E., MARS, M., MCWILLIAM, J., A. PLINT, D., PRAHL-ANDERSEN, B., SEMB, G. & C. SHAW, W. 1992. A Six-Center International Study of Treatment Outcome in Patients with Clefts of the Lip and Palate: Part 2. Craniofacial Form and Soft Tissue Profile. 29(5), 398-404

NESS, A. R., WILLS, A. K., WAYLEN, A., AL-GHATAM, R., JONES, T. E. M., PRESTON, R., IRELAND, A. J., PERSSON, M., SMALLRIDGE, J., HALL, A. J., SELL, D. & SANDY, J. R. 2015. Centralization of cleft care in the UK. Part 6: a tale of two studies. Orthod Craniofac Res, 18, 56-62.

NICHOLLS, W., SINGER, S. L., SOUTHALL, P. J. & WINTERS, J. C. 2014. The Assessment of Digital Study Models Using the GOSLON Yardstick Index. Cleft Palate Craniofac J, 51, 264-9.

NIGHTINGALE, C., WITHEROW, H., REID, F. D. & EDLER, R. 2003. Comparative reproducibility of three methods of radiographic assessment of alveolar bone grafting. Eur J Orthod, 25, 35-41.

NIRANJANE, P. P., KAMBLE, R. H., DIAGAVANE, S. P., SHRIVASTAV, S. S., BATRA, P., VASUDEVAN, S. D. & PATIL, P. 2014. Current status of presurgical infant orthopaedic treatment for cleft lip and palate patients: A critical review. Indian J Plast Surg. 47(03): 293-302

OHANNESSIAN, P., BERGGREN, A. & ABDIU, A. 2011. The cleft lip evaluation profile (CLEP): a new approach for postoperative nasolabial assessment in patients with unilateral cleft lip and palate. J Plast Surg Hand Surg, 45, 8-13.

OVSENIK, M. & PRIMOZIC, J. 2007. Evaluation of 3 occlusal indexes: Eismann index, Eismann-Farcnik index, and index of orthodontic treatment need. Am J Orthod Dentofacial Orthop, 131, 496-503.

PANGRAZIO-KULBERSH, V., KACZYNSKI, R. & SHUNOCK, M. 1999. Early treatment outcome assessed by the Peer Assessment Rating index. Am J Orthod Dentofacial Orthop, 115, 544-50.

PEGELOW, M., KLINTO, K., STALHAND, G., LEMBERGER, M., VESTERBACKA, M., RIZELL, S., CHALIEN, M. N., BJORNSTROM, L., BECKER, M., LINDBERG, M., MARCUSSON, A. & KARSTEN, A. 2020. Validation of reported dentoalveolar relationships in the Swedish Quality Registry for Cleft Lip and Palate. Eur J Orthod, 42, 30-35.

PERSSON, M., SANDY, J. R., WAYLEN, A., WILLS, A. K., AL-GHATAM, R., IRELAND, A. J., HALL, A. J., HOLLINGWORTH, W., JONES, T., PETERS, T. J., PRESTON, R., SELL, D., SMALLRIDGE, J., WORTHINGTON, H. & NESS, A. R. 2015. A cross-sectional survey of 5-year-old children with non-syndromic unilateral cleft lip and palate: the Cleft Care UK study. Part 1: background and methodology. Orthod Craniofac Res, 18 Suppl 2, 1-13.

PICARD, O. & WOOD, D. 2008. Medical Interviews: A Comprehensive Guide to CT, ST & Registrar Interview Skills : Over 120 Medical Interview Questions, Techniques & NHS Topics Explained, ISC Medical.

PITTS, N. B., EVANS, D. J. & PINE, C. M. 1997. British Association for the Study of Community Dentistry (BASCD) diagnostic criteria for caries prevalence surveys-1996/97. Community Dent Health, 14 Suppl 1, 6-9.

PUSIC, A. L., LEMAINE, V., KLASSEN, A. F., SCOTT, A. M. & CANO, S. J. 2011. Patient-reported outcome measures in plastic surgery: use and interpretation in evidence-based medicine. Plast Reconstr Surg, 127, 1361-7.

REAMES, B. N., GHAFERI, A. A., BIRKMEYER, J. D. & DIMICK, J. B. 2014. Hospital volume and operative mortality in the modern era. Ann Surg, 260, 244-51.

REVINGTON, P. J., MCNAMARA, C., MUKARRAM, S., PERERA, E., SHAH, H. V. & DEACON, S. A. 2010. Alveolar bone grafting: results of a national outcome study. Ann R Coll Surg Engl, 92, 643-6.

RICHMOND, S., SHAW, W. C., O'BRIEN, K. D., BUCHANAN, I. B., JONES, R., STEPHENS, C. D., ROBERTS, C. T. & ANDREWS, M. 1992. The development of the PAR Index (Peer Assessment Rating): reliability and validity. Eur J Orthod, 14, 125-39.

RICHMOND, S., TURBILL, E. A. & ANDREWS, M. 1993. Calibration of Non-dental and Dental Personnel in the Use of the PAR Index. British Journal of Orthodontics, 20, 231-234.

ROHRICH, R. J., LOVE, E. J., BYRD, H. S. & JOHNS, D. F. 2000. Optimal Timing of Cleft Palate Closure. Plastic and Reconstructive Surgery, 106(2):413-21.

ROSS, R. B. 1970. The clinical implications of facial growth in cleft lip and palate. Cleft Palate J, 7, 37-47.

RUSSELL, K., LONG, R. E., JR., DASKALOGIANNAKIS, J., MERCADO, A., HATHAWAY, R., SEMB, G. & SHAW, W. 2016. A Multicenter Study Using the SWAG Scale to Compare Secondary Alveolar Bone Graft Outcomes for Patients With Cleft Lip and Palate. Cleft Palate Craniofac J, 53, 180-6.

RUSSELL, K., LONG, R. E., JR., DASKALOGIANNAKIS, J., MERCADO, A., HATHAWAY, R., SEMB, G. & SHAW, W. 2017. Reliability of the SWAG-The Standardized Way to Assess Grafts Method for Alveolar Bone Grafting in Patients With Cleft Lip and Palate. Cleft Palate Craniofac J, 54, 680-686.

SANDY, J., RUMSEY, N., PERSSON, M., WAYLEN, A., KILPATRICK, N., IRELAND, T. & NESS, A. 2012. Using service rationalisation to build a research network: lessons from the centralisation of UK services for children with cleft lip and palate. Br Dent J, 212, 553-5.

SANDY, J., WILLIAMS, A., MILDINHALL, S., MURPHY, T., BEARN, D., SHAW, B., SELL, D., DEVLIN, B. & MURRAY, J. 1998. The Clinical Standards Advisory Group (CSAG) Cleft Lip and Palate Study. Br J Orthod, 25, 21-30.

SANDY, J. R., WILLIAMS, A. C., BEARN, D., MILDINHALL, S., MURPHY, T., SELL, D., MURRAY, J. J. & SHAW, W. C. 2001. Cleft lip and palate care in the United Kingdom--the Clinical Standards Advisory Group (CSAG) Study. Part 1: background and methodology. Cleft Palate Craniofac J, 38, 20-3.

SCALLY, G. & DONALDSON, L. J. 1998. Clinical governance and the drive for quality improvement in the new NHS in England. Br Med J, 317, 61-5.

SCHUSTER, C. 2004. A Note on the Interpretation of Weighted Kappa and its Relations to Other Rater Agreement Statistics for Metric Scales. Educational and Psychological Measurement, 64, 243-253.

SCHUTTE, B. C. & MURRAY, J. 1999. The many faces and factor of orofacial clefts. Hum Mol Genet. 8(10):1853-9.

SCOTT, J. K., LEARY, S. D., NESS, A. R., SANDY, J. R., PERSSON, M., KILPATRICK, N. & WAYLEN, A. E. 2014. Centralization of services for children born with orofacial clefts in the United kingdom: a cross-sectional survey. Cleft Palate Craniofac J, 51, e102-9.

SCOTT, J. K., LEARY, S. D., NESS, A. R., SANDY, J. R., PERSSON, M., KILPATRICK, N. & WAYLEN, A. E. 2015. Perceptions of team members working in cleft services in the United Kingdom: a pilot study. Cleft Palate Craniofac J, 52, e1-7.

SELL, D., GRUNWELL, P., MILDINHALL, S., MURPHY, T., CORNISH, T. A., BEARN, D., SHAW, W. C., MURRAY, J. J., WILLIAMS, A. C. & SANDY, J. R. 2001. Cleft lip and palate care in the United Kingdom--the Clinical Standards Advisory Group (CSAG) Study. Part 3: speech outcomes. Cleft Palate Craniofac J, 38, 30-7.

SELL, D., HARDING, A. & GRUNWELL, P. 1994. A screening assessment of cleft palate speech (Great Ormond Street Speech Assessment). Eur J Disord Commun, 29, 1-15.

SELL, D., HARDING, A. & GRUNWELL, P. 1999. GOS.SP.ASS.'98: an assessment for speech disorders associated with cleft palate and/or velopharyngeal dysfunction (revised). Int J Lang Commun Disord, 34, 17-33.

SELL, D., JOHN, A., HARDING-BELL, A., SWEENEY, T., HEGARTY, F. & FREEMAN, J. 2009. Cleft audit protocol for speech (CAPS-A): a comprehensive training package for speech analysis. Int J Lang Commun Disord, 44, 529-48.

SELL, D., MILDINHALL, S., ALBERY, L., WILLS, A. K., SANDY, J. R. & NESS, A. R. 2015. The Cleft Care UK study. Part 4: perceptual speech outcomes. Orthod Craniofac Res, 18 Suppl 2, 36-46.

SEMB, G., BRATTSTROM, V., MOLSTED, K., PRAHL-ANDERSEN, B. & SHAW, W. C. 2005. The Eurocleft study: intercenter study of treatment outcome in patients with complete cleft lip and palate. Part 1: introduction and treatment experience. Cleft Palate Craniofac J, 42, 64-8.

SHARMA, V. P., BELLA, H., CADIER, M. M., PIGOTT, R. W., GOODACRE, T. E. & RICHARD, B. M. 2012. Outcomes in facial aesthetics in cleft lip and palate surgery: a systematic review. J Plast Reconstr Aesthet Surg, 65, 1233-45.

SHAW, G. M. & LAMMER, E. J. 1999. Maternal periconceptional alcohol consumption and risk for orofacial clefts. J Pediatr, 134, 298-303.

SHAW, W. C., ASHER-MCDADE, C., BRATTSTROM, V., DAHL, E., MCWILLIAM, J., MOLSTED, K., PLINT, D. A., PRAHL-ANDERSEN, B., SEMB, G. & THE, R. P. 1992a. A six-center

international study of treatment outcome in patients with clefts of the lip and palate: Part 1. Principles and study design. Cleft Palate Craniofac J, 29, 393-7.

SHAW, W. C., DAHL, E., ASHER-MCDADE, C., BRATTSTROM, V., MARS, M., MCWILLIAM, J., MOLSTED, K., PLINT, D. A., PRAHL-ANDERSEN, B. & ROBERTS C. 1992b. A six-center international study of treatment outcome in patients with clefts of the lip and palate: Part 5. General discussion and conclusions. Cleft Palate Craniofac J, 29, 413-8.

SHAW, W. C., MEEK, S. C. & JONES, D. S. 1980. Nicknames, teasing, harassment and the salience of dental features among school children. Br J Orthod, 7, 75-80.

SHAW, W. C., WILLIAMS, A. C., SANDY, J. R. & DEVLIN, H. B. 1996. Minimum standards for the management of cleft lip and palate: efforts to close the audit loop. Royal College of Surgeons of England. Ann R Coll Surg Engl, 78, 110-4.

SHI, B. & LOSEE, J. E. 2014. The impact of cleft lip and palate repair on maxillofacial growth. International Journal of Oral Science, 7, 14-17.

SIM, J. & WRIGHT, C. C. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther, 85, 257-68.

SMALLRIDGE, J., HALL, A. J., CHORBACHI, R., PARFECT, V., PERSSON, M., IRELAND, A. J., WILLS, A. K., NESS, A. R. & SANDY, J. R. 2015. Functional outcomes in the Cleft Care UK study – Part 3: oral health and audiology. Orthod Craniofac Res. 18(Suppl Suppl 2): 25–35.

SOUTHALL, P., WALTERS, M. & SINGER, S. 2012. The influence of orthodontic treatment on the Goslon score of unilateral cleft lip and palate patients. Cleft Palate Craniofac J, 49, 215-20.

SPRIESTERSBACH, D. C., DICKSON, D. R., FRASER, F. C., HOROWITZ, S. L., MCWILLIAMS, B. J., PARADISE, J. L. & RANDALL, P. 1973. Clinical research in cleft lip and cleft palate: the state of the art. Cleft Palate J, 10, 113-65.

STEINER, D. L., NORMAN, G. R. & CAIRNEY, J. 2008. Health Measurement Scales: A practical guide to their development and use, Oxford, UK, Oxford University Press.

TAGHAVI BAYAT, J., HUGGARE, J., MOHLIN, B. & AKRAMI, N. 2017. Predicting orthodontic treatment need: reliability and validity of the Demand for Orthodontic Treatment Questionnaire. Eur J Orthod, 39, 326-333.

TAN, E. L. Y., KUEK, M. C., WONG, H. C., ONG, S. A. K. & YOW, M. 2018. Secondary Dentition Characteristics in Children With Nonsyndromic Unilateral Cleft Lip and Palate: A Retrospective Study. Cleft Palate Craniofac J, 55, 582-589.

TOLLEFSON, T. T., WHITE, D., BROOKES, J. & GOUDY, S. 2012. Velopharyngeal Insufficiency and Cleft. Int J Otolaryngol, 2012: 864069.

TOTHILL, C. & MOSSEY, P. A. 2007. Assessment of arch constriction in patients with bilateral cleft lip and palate and isolated cleft palate: a pilot study. Eur J Orthod, 29, 193-7.

TSANGARIS, E., WONG RIFF, K. W., GOODACRE, T., FORREST, C. R., DREISE, M., SYKES, J., DE CHALAIN, T., HARMAN, K., O'MAHONY, A., PUSIC, A. L., THABANE, L., THOMA, A. & KLASSEN, A. F. 2017. Establishing Content Validity of the CLEFT-Q: A New Patient-reported Outcome Instrument for Cleft Lip/Palate. Plast Reconstr Surg Glob Open, 5(4): e1305.

TURNER, S. R., RUMSEY, N. & SANDY, J. R. 1998. Psychological aspects of cleft lip and palate. Eur J Orthod, 20, 407-15.

UZEL, A. & ALPARSLAN, Z. N. 2011. Long-Term Effects of Presurgical Infant Orthopedics in Patients with Cleft Lip and Palate: A Systematic Review. Cleft Palate Craniofac J, 48, 587-595.

WAITE, P. D. & WAITE, D. E. 1996. Bone grafting for the alveolar cleft defect. Semin Orthod, 2, 192-6.

WAYLEN, A., NESS, A. R., WILLS, A. K., PERSSON, M., RUMSEY, N. & SANDY, J. R. 2015. Cleft Care UK study. Part 5: child psychosocial outcomes and satisfaction with cleft services. Orthod Craniofac Res, 18 Suppl 2, 47-55.

WERLER, M. M., LAMMER, E. J., ROSENBERG, L. & MITCHELL, A. A. 1991. Maternal alcohol use in relation to selected birth defects. Am J Epidemiol, 134, 691-8.

WILLADSEN, E., BOERS, M., SCHOPS, A., KISLING-MOLLER, M., NIELSEN, J. B., JORGENSEN, L. D., ANDERSEN, M., BOLUND, S. & ANDERSEN, H. S. 2018. Influence of timing of delayed hard palate closure on articulation skills in 3-year-old Danish children with unilateral cleft lip and palate. Int J Lang Commun Disord, 53, 130-143.

WILLADSEN, E., LOHMANDER, A., PERSSON, C., LUNDEBORG, I., ALALUUSUA, S., AUKNER, R., BAU, A., BOERS, M., BOWDEN, M., DAVIES, J., EMBORG, B., HAVSTAM, C., HAYDEN, C., HENNINGSSON, G., HOLMEFJORD, A., HOLTTA, E., KISLING-MOLLER, M., KJOLL, L., LUNDBERG, M., MCALEER, E., NYBERG, J., PAASO, M., PEDERSEN, N. H., RASMUSSEN, T., REISAETER, S., ANDERSEN, H. S., SCHOPS, A., TORDAL, I. B. & SEMB, G. 2017. Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 5. Speech

outcomes in 5-year-olds - consonant proficiency and errors. J Plast Surg Hand Surg, 51, 38-51.

WILLIAMS, A., SHAW, W. C. & DEVLIN, H. B. 1994. Provision of services for cleft lip and palate in England and Wales. Br Med J, 309, 1552.

WILLIAMS, A. C., BOWER, E. J. & NEWTON, J. T. 2004. Research in primary dental care part 4: measures. Br Dent J, 196, 739-46.

WILLIAMS, A. C., SHAW, W. C., SANDY, J. R. & DEVLIN, H. B. 1996. The surgical care of cleft lip and palate patients in England and Wales. Br J Plast Surg, 49, 150-5.

WITHEROW, H., COX, S., JONES, E., CARR, R. & WATERHOUSE, N. 2002. A new scale to assess radiographic success of secondary alveolar bone grafts. Cleft Palate Craniofac J, 39, 255-60.

ZAPF, A., CASTELL, S., MORAWIETZ, L. & KARCH, A. 2016. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? BMC Medical Research Methodology, 16, 93.

ZHU, S., YANG, Y., GU, M. & KHAMBAY, B. 2016. A Comparison of Three Viewing Media for Assessing Dental Arch Relationships in Patients With Unilateral Cleft Lip and Palate. Cleft Palate Craniofac J, 53, 578-83.

APPENDIX 1: INFORMATION SHEET ON USE OF THE MODIFIED 5-YEAR-OLDS' INDEX FOR GROUP 1

How to use the Modified 5-Year-Olds' Index

The Modified 5-Year-Olds' Index was developed by Mittal *et al.* (2018) in order to increase the discriminatory power of the original 5-Year-Olds' Index in assessing the dentoalveolar outcome of primary cleft surgery at the age of 5. Categories 2 & 3 of the original 5-category index were expanded into four categories, resulting in a 7-category index. The verbal descriptors of the Modified 5-Year-Olds' index aid categorisation of models. Category 1 represents the best possible outcome, with Category 7 representing the worst. The table below shows the original 5-Year-Olds' Index alongside the Modified Index, demonstrating the five categories created from the original categories 1-3.

5-Year Olds' Category	Modified Category	Features
I	I	Good positive overjet Good positive overbite Good archform Class II or I dentoalveolar relationship
2	2	Good positive overjet Crossbite on C only Class II/2 or Class I incisors
	3	Positive overjet Crossbite on some teeth in lesser segment (but some teeth not) Edge-to-edge incisors with no crossbites
3	4	Class III incisors Reducing overbite Nearly complete unilateral crossbite
	5	Edge-to-edge incisors Reduced/tenuous overbite Marked dentoalveolar compensation Unilateral crossbite
4	6	Negative overjet, incisors may be contacting Lower arch compensation Bilateral crossbite tendency Anterior open bite developing
5	7	Large reverse overjet Bilateral crossbite Anterior open bite

The 50 models for assessment have been randomly selected from the CCUK study, representing a range of dental arch relationships in UCLP cases. A subjective assessment of the dentoalveolar features is made using the verbal descriptors of the index for reference, and a category subsequently chosen.

In assessment of models, certain features are considered most important:

- Overjet
- AP relationship

The position of the incisors can be mentally decompensated in order to visualise the AP relationship. If teeth are missing, their probable position within the alveolus should be visualised during assessment. The transverse relationship is less important than the overjet and AP relationship. Vertical defects are not considered important as these defects are addressed with alveolar bone grafting at a later stage.

The majority of study models are easily categorised within a minute. Use of the Index is subjective and judgement is needed, particularly in assessment of the more difficult cases.

References

Mittal, TK, Ireland, AJ, Atack, NE, Leary, SD, Russell, JI, Deacon, SA, Ness, AR & Sandy, JR, 2018, 'Outcome measures in ULCP: the modified 5-Year-Olds'-Index-development and reliability'. *Cleft Palate-Craniofacial Journal,* 56, 248-256.

The Modified 5-Year-Olds' Index

Category	Features
1	Good positive overjet Good positive overbite Good archform Class II or I dentoalveolar
2	Good positive overjet Crossbite on C only Class II/2 or Class I incisors
3	Positive overjet Crossbite on some teeth in lesser segment (but some teeth not) Edge to Edge incisors with no crossbites
4	Class III incisors (positive overjet) Reducing overbite Nearly complete unilateral crossbite
5	Edge to Edge incisors Reduced/tenuous overbite Marked dentoalveolar compensation Unilateral crossbite
6	Negative overjet, incisors may be contacting Lower arch compensation Bilateral crossbite tendency Anterior openbite devleoping
7	Large reverse overjet Bilateral crossbite Anterior openbite

APPENDIX 2: INFORMATION SHEET ON USE OF THE MODIFIED 5-YEAR-OLDS' INDEX FOR GROUPS 2 & 3

How to use the Modified 5-Year-Olds' Index

The Modified 5-Year-Olds' Index was developed by Mittal *et al.* (2018) in order to increase the discriminatory power of the original 5-Year-Olds' Index in assessing the dentoalveolar outcome of primary cleft surgery at the age of 5. Categories 2 & 3 of the original 5-category index were expanded into four categories, resulting in a 7-category index. The verbal descriptors of the Modified 5-Year-Olds' index aid categorisation of models. Category 1 represents the best possible outcome, with Category 7 representing the worst. The table below shows the original 5-Year-Olds' Index alongside the Modified Index, demonstrating the five categories created from the original categories 1-3.

5-Year Olds' Category	Modified Category	Features
I	I	Good positive overjet Good positive overbite Good archform Class II or I dentoalveolar relationship
2	2	Good positive overjet Crossbite on C only Class II/2 or Class I incisors
	3	Positive overjet Crossbite on some teeth in lesser segment (but some teeth not) Edge-to-edge incisors with no crossbites
3	4	Class III incisors Reducing overbite Nearly complete unilateral crossbite
	5	Edge-to-edge incisors Reduced/tenuous overbite Marked dentoalveolar compensation Unilateral crossbite
4	6	Negative overjet, incisors may be contacting Lower arch compensation Bilateral crossbite tendency Anterior open bite developing
5	7	Large reverse overjet Bilateral crossbite Anterior open bite

The 50 models for assessment have been randomly selected from the CCUK study, representing a range of dental arch relationships in UCLP cases. A subjective assessment of the dentoalveolar features is made using the verbal descriptors of the index for reference, and a category subsequently chosen.

In assessment of models, certain features are considered most important:

- Overjet
- AP relationship

The position of the incisors can be mentally decompensated in order to visualise the AP relationship. If teeth are missing, their probable position within the alveolus should be visualised during assessment. The transverse relationship is less important than the overjet and AP relationship. Vertical defects are not considered important as these defects are addressed with alveolar bone grafting at a later stage.

The majority of study models are easily categorised within a minute. Use of the Index is subjective and judgement is needed, particularly in assessment of the more difficult cases.

A set of 14 reference models has been provided (two per category), representing examples of features expected. As the models are only examples of features that may present in each category, it is important that there is no attempt to 'match' these models to the study models being assessed.

References

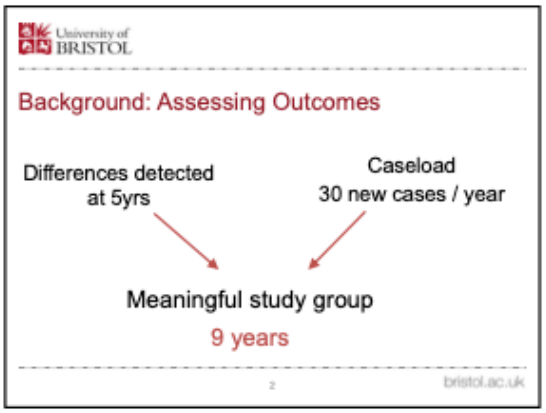
Mittal, TK, Ireland, AJ, Atack, NE, Leary, SD, Russell, JI, Deacon, SA, Ness, AR & Sandy, JR, 2018, 'Outcome measures in ULCP: the modified 5-Year-Olds'-Index-development and reliability'. *Cleft Palate-Craniofacial Journal,* 56, 248-256.

The Modified 5-Year-Olds' Index

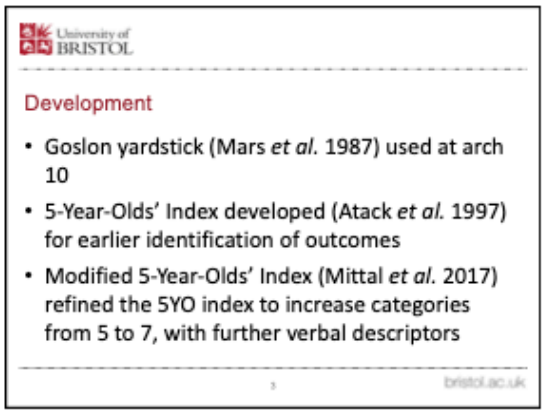
Category	Features
1	Good positive overjet Good positive overbite Good archform Class II or I dentoalveolar
2	Good positive overjet Crossbite on C only Class II/2 or Class I incisors
3	Positive overjet Crossbite on some teeth in lesser segment (but some teeth not) Edge to Edge incisors with no crossbites
4	Class III incisors (positive overjet) Reducing overbite Nearly complete unilateral crossbite
5	Edge to Edge incisors Reduced/tenuous overbite Marked dentoalveolar compensation Unilateral crossbite
6	Negative overjet, incisors may be contacting Lower arch compensation Bilateral crossbite tendency Anterior openbite devleoping
7	Large reverse overjet Bilateral crossbite Anterior openbite

APPENDIX 3: MODIFIED 5-YEAR-OLDS' INDEX CALIBRATION COURSE

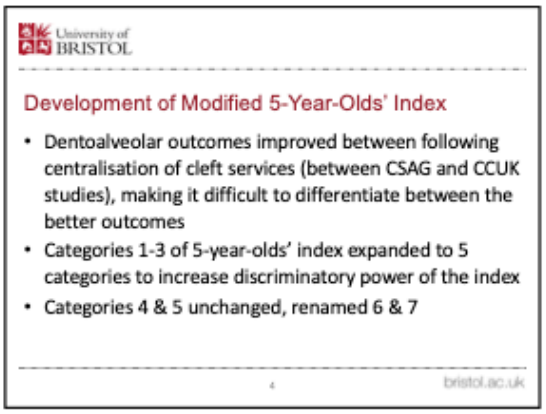
1



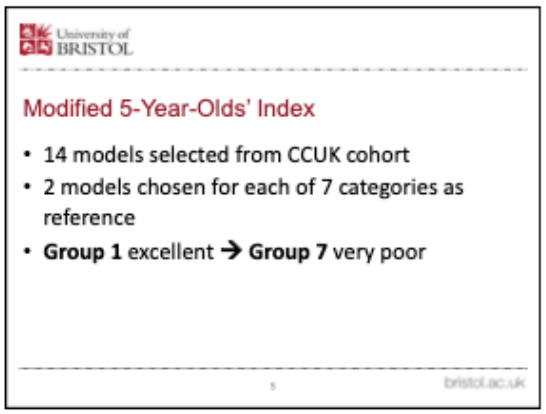
2



3



4



5

6

University of BRISTOL

Modified 5-Year-Olds' Index

- Overall subjective assessment
- 7 categories
  - Group 1 – excellent
  - Group 7 – very poor
- 2 reference models per category
- Represents the full range of clinical presentation

7 bristol.ac.uk

7

University of BRISTOL

Use of Index

- Most important features
  - Overjet
  - AP relationship
- Less important features
  - Transverse relationship
- Least important features
  - Vertical defects around cleft site

8 bristol.ac.uk

8

University of BRISTOL

Use of Index

- Assessing AP relationship
  - Mentally decompensate incisors
  - Visualise probable incisor position within the alveolus if these teeth are missing
- Vertical defects
  - Least important due to improved success rates of secondary alveolar bone grafting

9 bristol.ac.uk

9

University of BRISTOL

Major Features of Modified 5-Year-Olds' Index

10 bristol.ac.uk

10

University of BRISTOL

Category 1

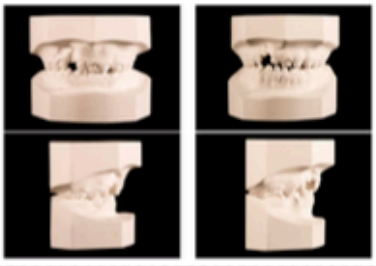
- Features
  - Good positive overjet
  - Good positive overbite
  - Good archform
  - Class II or I dentoalveolar relationship

11 bristol.ac.uk

11

University of BRISTOL

Category 1 reference models



12 bristol.ac.uk

12

University of BRISTOL

Category 2

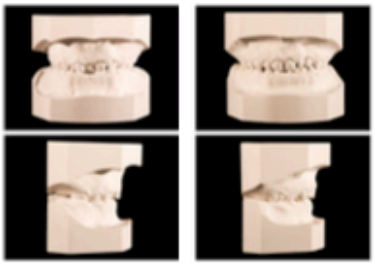
- Features
  - Good positive overjet
  - Crossbite on C only
  - Class II/2 or Class I incisors

13 bristol.ac.uk

13

University of BRISTOL

Category 2 reference models



14 bristol.ac.uk

14

University of BRISTOL

Category 3

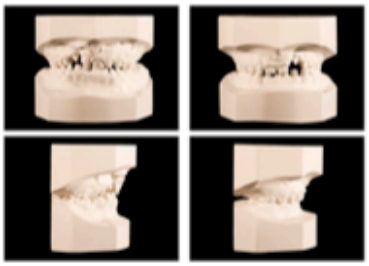
- Features
  - Positive overjet
  - Crossbite on some teeth in lesser segment (but some teeth not)
  - Edge-to-edge incisors with no crossbites

15 bristol.ac.uk

15

University of BRISTOL

Category 3 reference models



16 bristol.ac.uk

16

University of BRISTOL

Category 4

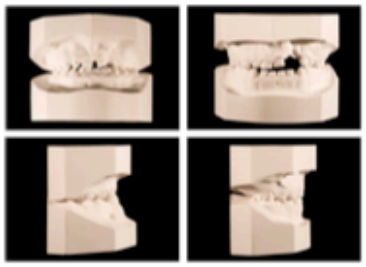
- Features
  - Class III incisors
  - Reducing overbite
  - Nearly complete unilateral crossbite

17 bristol.ac.uk

17

University of BRISTOL

Category 4 reference models



18 bristol.ac.uk

18

University of BRISTOL

Category 5

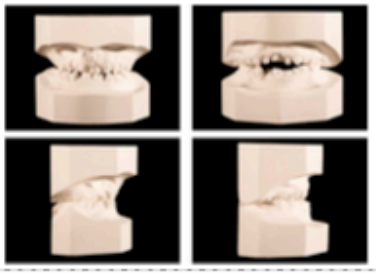
- Features
  - Edge-to-edge incisors
  - Reduced/tenuous overbite
  - Marked dentoalveolar compensation
  - Unilateral crossbite

19 bristol.ac.uk

19

University of BRISTOL

Category 5 reference models



20 bristol.ac.uk

20

University of BRISTOL

Category 6

- Features
  - Negative overjet, incisors may be contacting
  - Lower arch compensation
  - Bilateral crossbite tendency
  - Anterior openbite developing

21 bristol.ac.uk

21

University of BRISTOL

Category 6 reference models

22 bristol.ac.uk

22

University of BRISTOL

Category 7

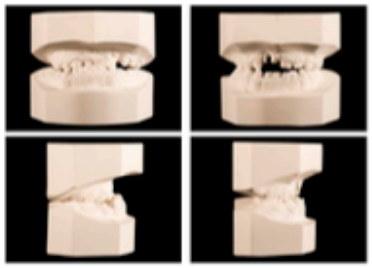
- Features
  - Large reverse overjet
  - Bilateral crossbite
  - Anterior openbite

23 bristol.ac.uk

23

University of BRISTOL

Category 7 reference models



24 bristol.ac.uk

24

University of BRISTOL

How to assess a model

25

University of BRISTOL

Summary

- 'Scoring order' only guidance
- Use reference models as an aid
- Subjective assessment
  - Borderline cases can be difficult
- Limitations
  - Assessment using models only
  - UCLP only
  - Missing teeth & AOB not accounted for
  - Developed for Caucasian population

26

APPENDIX 4: SCORE SHEET

Modified-5-Year-Olds' Index Scoring Sheet

Name…… Session……

Model number	M5YO score
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
21	
22	
23	
24	
25	

Model number	M5YO score
26	
27	
28	
29	
30	
31	
32	
33	
34	
35	
36	
37	
38	
39	
40	
41	
42	
43	
44	
45	
46	
47	
48	
49	
50	

APPENDIX 5: CROSS-TABULATED OF SCORES BETWEEN ASSESSORS AND GOLD STANDARD EXPERT CONSENSUS MODIFIED 5-YEAR-OLDS' INDEX SCORE FOR SESSION 2

Group 1 Session 2

Gold standard consensus score	Assessor 1 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2	1						3
	2		3	1					4
	3		1	6	5				12
	4				5	3			8
	5				2	7	1		10
	6					2	7		9
	7						2	2	4
	Total	2	5	7	12	12	10	2	50

Table 40: Cross tabulation of Assessor 1 (consultant) scores for session 2 and gold standard expert consensus score

Gold standard consensus score	Assessor 2 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	1	1	1					3
	2		2	2					4
	3		1	8	3				12
	4				5	2	1		8
	5				2	7	1		10
	6					3	6		9
	7						2	2	4
	Total	1	4	11	10	12	10	2	50

Table 41: Cross tabulation of Assessor 2 (post-CCST trainee) scores for session 2 and gold standard expert consensus score

Gold standard consensus score	Assessor 3 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2	1						3
	2		2	2					4
	3		1	6	3	1	1		12
	4			3	1	3		1	8
	5			2	1	4	3		10
	6					1	8		9
	7						3	1	4
	Total	2	4	13	5	9	15	2	50

Table 42: Cross tabulation of Assessor 3 (specialty trainee) scores for session 2 and gold standard expert consensus score

Gold standard consensus score	Assessor 4 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	1	1	1					3
	2		2	2					4
	3			3	4	4			11
	4				6	2	1		9
	5			1	4	4			9
	6				1	2	1		4
	7					1	6	3	10
	Total	1	3	7	15	13	8	3	50

Table 43: Cross tabulation of Assessor 4 (specialty trainee) scores for session 2 and gold standard expert consensus score

Gold standard consensus score	Assessor 5 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	1	1	1					3
	2		1	2	1				4
	3		3	5	2	2			12
	4			2	3	2		1	8
	5				1	3	6		10
	6					2	6	1	9
	7							4	4
	Total	1	5	10	7	9	12	6	50

Table 44: Cross tabulation of Assessor 5 (specialty trainee) scores for session 2 and gold standard expert consensus score

Group 2 session 2

Gold standard consensus score	Assessor 6 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		3	1					4
	3		2	3	7				12
	4				5	3			8
	5				3	5	2		10
	6					2	6	1	9
	7							4	4
	Total	3	5	4	15	10	8	5	50

Table 45: Cross tabulation of Assessor 6 (consultant) scores for session 2 and gold standard expert consensus score

Gold standard consensus score	Assessor 7 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		3	1					4
	3		5	5	2				12
	4			3	5				8
	5				3	6	1		10
	6				1	4	4		9
	7						3	1	4
	Total	3	8	9	11	10	8	1	50

Table 46: Cross tabulation of Assessor 7 (post-CCST trainee) scores for session 2 and gold standard expert consensus score

Gold standard consensus score	Assessor 8 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		3	1					4
	3		2	8	2				12
	4		1	2	2	2	1		8
	5			1	1	8			10
	6				1	1	7		9
	7					1	2	1	4
	Total	3	6	12	6	12	10	1	50

Table 47: Cross tabulation of Assessor 8 (specialty trainee) scores for session 2 and gold standard expert consensus score

Gold standard consensus score	Assessor 9 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		3	1					4
	3		1	9	2				12
	4			6	1	1			8
	5			1	4	4	1		10
	6				1	1	7		9
	7				1		1	2	4
	Total	3	4	17	9	6	9	2	50

Table 48: Cross tabulation of Assessor 9 (specialty trainee) scores for session 2 and gold standard expert consensus score

Gold standard consensus score	Assessor 10 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		3	1					4
	3			8	3		1		12
	4			2	4	2			8
	5			2	1	5	2		10
	6					2	7		9
	7							4	4
	Total	3	3	13	8	9	10	4	50

Table 49: Cross tabulation of Assessor 10 (specialty trainee) scores for session 2 and gold standard expert consensus score

Group 3 Session 2

Gold standard consensus score	Assessor 11 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		2	1	1				4
	3	1		6	4		1		12
	4			2	6				8
	5				3	6	1		10
	6					5	4		9
	7				1		2	1	4
	Total	4	2	9	15	11	8	1	50

Table 50: Cross tabulation of Assessor 11 (consultant) scores for session 2 and gold standard expert consensus score

Gold standard consensus score	Assessor 12 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		3	1					4
	3			11	1				12
	4			5	3				8
	5			2	2	4	2		10
	6					5	4		9
	7						2	2	4
	Total	3	3	19	6	9	8	2	50

Table 51: Cross tabulation of Assessor 12 (post-CCST trainee) scores for session 2 and gold standard expert consensus score

Gold standard consensus score	Assessor 13 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	3							3
	2		1	2	1				4
	3			5	7				12
	4				2	6			8
	5					9	1		10
	6					3	6		9
	7						2	2	4
	Total	3	1	7	10	18	9	2	50

Table 52: Cross tabulation of Assessor 13 (specialty trainee) scores for session 2 and gold standard expert consensus score

Gold standard consensus score	Assessor 14 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	2	1						3
	2		1	3					4
	3		1	9	2				12
	4			4	2	2			8
	5			1	6	3			10
	6				2	4	3		9
	7				1		2	1	4
	Total	2	3	17	13	9	5	1	50

Table 53: Cross tabulation of Assessor 14 (specialty trainee) scores for session 2 and gold standard expert consensus score

Gold standard consensus score	Assessor 15 – Session 2								
	M5YO Index Score	1	2	3	4	5	6	7	**Total**
	1	1	2						3
	2		2	2					4
	3		1	5	6				12
	4			3	3	2			8
	5					7	3		10
	6					5	4		9
	7						2	2	4
	Total	1	5	10	9	14	9	2	50

Table 54: Cross tabulation of Assessor 15 (specialty trainee) scores for session 2 and gold standard expert consensus score